ADDRESSING DATA INFORMATIVENESS IN RISK-CONSCIOUS BUILDING PERFORMANCE SIMULATION APPLICATIONS

A Dissertation
Presented to
The Academic Faculty

by

Qi Li

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Architecture

Georgia Institute of Technology
August 2017

ADDRESSING DATA INFORMATIVENESS IN RISK-CONSCIOUS BUILDING PERFORMANCE SIMULATION APPLICATIONS

Approved by:

Dr. Jason Brown, Advisor
School of Architecture
Georgia Institute of Technology

Dr. Ruchi Choudhary
Department of Engineering
University of Cambridge

Prof. Godfried Augenbroe, Co-Advisor
School of Architecture
Georgia Institute of Technology

Ron Judkoff
Buildings and Thermal Systems Center
National Renewable Energy Laboratory

Dr. Ralph Muehleisen
Energy Systems Division
Argonne National Laboratory

Dr. C. F. Jeff Wu
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: July 20, 2017

To my family for their continuous love and support

ACKNOWLEDGEMENTS

This work would not have been accomplished without the help of many people in my five years' Ph.D. study. First and foremost, I would like to thank my research advisor Dr. Jason Brown for his tremendous guidance and encouragement. I'm deeply indebted to him for giving me the freedom to pursue my research interest, and for the excellent example he has provided as a meticulous and passionate researcher.

I'm sincerely grateful to the privilege to have Prof. Godfried Augenbroe as my co-advisor, for his profound knowledge and sharp insights in nurturing my research perspectives and immense support to both of my academic and life pursuits. His word "explore new options" brought me to Georgia Tech and opened the door to a new world of research for me.

I would like to thank Dr. Ralph Muehleisen for the opportunity to participate in a set of inspirational projects, and his patient review of this dissertation and constructive comments. I also wish to thank Dr. Ruchi Choudhary for her enlightening advice and endless support during my internship at University of Cambridge, a wonderful research and life experience that contributes significantly to this work. I would like to extend my gratitude to Ron Judkoff and Dr. C.F. Jeff Wu for serving on my dissertation committee and sharing their valuable comments and suggestions. A special note of appreciation goes to Dr. Yeonsook Heo for her significant contributions to this work and considerate help to my life in Cambridge.

My Ph.D. study has greatly benefited from diverse research experiences. Special thanks to Yuming Sun and Qinpeng Wang for their extensive knowledge that shaped my graduate research, and to Dr. Perry Yang, Steven Jige Quan, Dr. Kathrin Menberg, Rebecca Ward, Li Gu, Dr. John Haymaker, and Dr. Benjamin Welle for their guidance and help during our collaborations. I also wish to thank my fellow graduate students Yuna Zhang, Ji-Hyun Kim, Roya Rezaee, Atefe Makhmalbaf, Sang-Hoon Lee, Michael Street, Gustavo Carneiro, Yiyuan Jia, Yifu Shi, Di Lu, Zhaoyun Zeng, and my friends Te, Di, Zengfeng, Jaesuk, Natalia, Ziping, Chen, Junfan, and Tian, without whom my life at Atlanta would not be so wonderful and memorable.

A heartfelt appreciation goes to my dearest Tianyao, who has always stood by me with her continuous encouragement and unreserved support throughout this journey. Last but not least, I would like to express my deepest appreciation to my parents and my sister. I could not become the person I am today without their considerate understanding, selfless sacrifice, and unconditional love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AHU	Air handling unit
BMS	Building management system
BPS	Building performance simulation
CRPS	Continuous rank probability score
CV-RMSE	Coefficient of variation of root mean square error
ECM	Energy conservation measure
HMC	Hamiltonian Monte Carlo
HVAC	Heating, ventilation, and air-conditioning
LHD	Latin Hypercube design
MCMC	Markov chain Monte Carlo
MCRPS	Mean continuous rank probability score
nMCRPS	Normalized mean continuous rank probability score
nMAE	Normalized mean absolute error
NMBE	Normalized mean bias error
RMSE	Root mean square error
TMY	Typical meteorology year
TRV	Thermostatic radiator valve

SUMMARY

Building performance management remains an important aspect in reducing building energy consumption and enhancing occupants' thermal comfort and productivity. Recent decades witnessed the maturity and proliferation of numerous methods, software and tools that span the whole spectrum of common building performance management practice. Among those related research and applications, the use of information and data in calibration and validation of building performance simulation (BPS) models constitutes an important subject of study especially in fault detection, operations management, and retrofit analysis. An extensive review of BPS model calibration and validation studies reveals two major research gaps. First, contemporary model calibration practice calls for an effective and robust method that can systematically incorporate a variety of information and data, handle modelling and prediction uncertainties, and maintain consistent model accuracy. Second, current approaches to collecting information and data in real applications largely depend on individual experience or common practice; further study is needed to understand the value of information and data, i.e. assess data informativeness, such as to support specific decision-making processes in choosing data monitoring strategies and to avoid missed opportunities or wasted resources.

To this end, this dissertation develops a new framework to address data informativeness in model calibration and validation to answer two major research questions: 1) how to make optimal use of available information and data to calibrate a building simulation model under uncertainty, and 2) how to quantify the informativeness of information and data for risk-conscious building performance simulation applications.

This framework builds upon uncertainty propagation using detailed measurements, and inverse modelling using Bayesian inference. It also introduces probabilistic accuracy metrics to assess model prediction accuracy, and uses explicit risk assessment to quantify data informativeness. Following an explanation of the framework's theoretical soundness, this dissertation provides two case studies to demonstrate its practical effectiveness. The first is a controlled experiment in the FLEXLAB test facility at Lawrence Berkeley National Laboratory. A new validation methodology is proposed to validate a simulation model under uncertainty, in which the validation criteria build upon the introduced probabilistic accuracy metrics. Given the experiment setup, uncertainty propagation based on synthetic measurements is applied, which effectively improves prediction agreement and reduces the risk of accepting invalid simulation outcomes. The second is to determine the appropriate model form and metering data for a hypothetical intervention analysis of an existing building with hydronic heating on the Cambridge, UK campus. A three-level modelling method is proposed to enable modelling all the thermal processes occurring in individual rooms while efficiently modelling the whole building to estimate heating system performance. Different sets of metering data are then used to calibrate the physical model, and the result indicates the superiority of Bayesian inference in exploiting the value of data, the necessity of electricity monitoring under uncontrolled conditions, and the potential of daily metering data for calibration in real building performance management practice.

CHAPTER 1. INTRODUCTION

Building performance management remains an important aspect in reducing building energy consumption and associated operation cost, and enhancing occupants' indoor thermal environment and productivity. In 2016, residential and commercial buildings consumed about 39% of total U.S. energy consumption (EIA, 2017). The Paris Carbon Agreement imposes new challenges for world- and nation-wide carbon emission mitigations, and calls for further research and development to explore innovative solutions in real practice.

Building performance management has continuously benefited from advances in building performance simulation (BPS) applications. Recent decades witnessed the maturity and proliferation of numerous methods, software and tools that span the whole spectrum of common applications. Among those related research and applications, the use of information and data in calibration and validation of BPS models constitutes an important subject of study especially in fault detection, operations management, and retrofit analysis. The growing availability of extensive operations data, thanks to the rapid progress in sensor and monitoring technology as well as the development of smart building and Internet of Things, provides great opportunities for a new era of BPS research and development. This would include continuous investigation of comprehensive and efficient modelling methods, systematic study of model performance within a risk-conscious decision-making context, and further exploration of effective use of information and data. This dissertation attempts to address data informativeness in the context of model

calibration and validation, and starts with a survey of the historical research in the related literature.

1.1 Literature review

A brief yet extensive literature review of current BPS model calibration and validation methods is presented in this section. Similar work on calibration methods include those by Reddy (2006), Coakley et al. (2014), and partially by Chaudhary et al. (2016), whereas a partial counterpart in model validation is presented by Judkoff and Neymark (2006). Ma et al. (2012) presented a summary that covers a wide range of topics in building energy retrofit that includes model calibration. The review of Fumo (2014) focuses on general methods of building energy estimation where model calibration constitutes an important aspect. In particular, Fabrizio and Monetti (2015) presented an insightful summary of the state of the art in calibration literature following a similar classification logic. This section distinguishes itself from the above work by focusing on the handling of model uncertainties and the collection and use of information and data in the literature review and identification of research gaps.

1.1.1 Calibration method

A proliferation of research on forward uncertainty propagation exists in the literature. Related summaries can be found in the work of Yao et al. (2011) and Wang (2016). This section focuses on inverse modelling methods in which the observations directly inform the calibration process.

1.1.1.1 Manual model tuning

The most common calibration method in real practice is manual model tuning, i.e. modeler changes certain model parameter values in a heuristic and iterative manner to reconcile predictions with observations. Most studies in this category focus on the procedures, guidance, and particularly graphical and analytical techniques that inform the model tuning process, which otherwise would be solely based on modeler's knowledge and experience.

In addition to conventional monthly and hourly time series plot, advanced graphical techniques have been proposed and applied in the literature. This includes 3D plots (Bronson et al., 1992; Haberl and Bou-Saada, 1998), color contours (Haberl et al., 1996; Raftery and Keane, 2011), and binned box-whisker mean plots (Haberl and Bou-Saada, 1998). These visualization techniques in principle provide a convenient and comprehensive overview of the model discrepancy to the modeler, who can then easily identify the underlying patterns and continue to the next iteration.

Another type of facilitated manual model tuning involves parametric and sensitivity analyses (Coakley et al., 2011; O'Neill et al., 2011; Reddy et al., 2007a; Westphal and Lamberts, 2005). Instead of iterative parameter value adjustment, this method performs parametric runs of the model a priori to identify influential parameters, and possible combinations of parameter values that reconcile model predictions with the observations.

Some studies focused on the direct investigation of prediction residuals, i.e. the deviation between deterministic model predictions and their respective observations. The underlying logic is that by pooling sufficient data together, those residuals should distribute randomly, i.e. exhibit no explicit pattern with respect to any variables, if the model can

explain the physical process well. Sun et al. (2016) proposed a pattern-recognition based model calibration method that follows this logic, but as this method solely uses monthly utility bills, insufficient data points may impair the confidence of calibration result. Palomo et al. (1991) developed a more comprehensive and statistically rigorous method of residual analysis, and Clarke et al. (1993) applied this method to calibrate a simulation model of a test cell. The use of hourly high quality data presumably increases the model's validity, but the cost and effort associated with extensive monitoring prohibits its application in an actual building with larger complexity and variability.

In the meantime, signature analysis was proposed, formalized, and extended in a series of studies (Liu and Liu, 2011; Liu et al., 2004; Liu and Claridge, 1998) and constitutes the ASHRAE 1092-RP research project (Liu et al., 2006). This method uses parametric and graphical techniques to facilitate residual analysis in a systematic manner. This method involves visual inspection of both the calibration signature, i.e. the normalized residuals between measured energy consumption and the corresponding simulated values as a function of outdoor air temperature, and characteristic signature, i.e. the counterpart in simulation results obtained by varying parameter values sequentially from default value. Comparison between these two types of energy signature helps the modeler to identify possible cause of discrepancy from incorrect parameter values.

In addition, evidence-based model calibration approaches have been proposed (Coakley et al., 2011; Monfet et al., 2009; Raftery et al., 2011). These approaches often involve explicit representation and documentation of human knowledge, which includes information and data source hierarchy and record of decisions (Raftery et al., 2011), residual patterns (Sun et al., 2016), calibration procedures (Pan et al., 2007; Yoon et al.,

2003), etc. This practice improves the reliability and reproducibility of manual calibration process.

To summarize, albeit facilitated by advanced analytical and graphical techniques, manual model tuning still requires human interventions to select and tune model parameters. Although these interventions can be standardized into knowledge-based expert rules, human subjectivity remains employed implicitly in guiding the calibration process in most cases. With the presence of good quality of information and data and very experienced modeler, manual calibration often results in models of high quality and credibility. Nevertheless, this approach in general is difficult to maintain reproducibility in real practices, usually demands significant time and labor, and may be vulnerable to risks in extreme cases where previous knowledge and experience do not apply. In the meantime, its incapability to systematically handle model uncertainty and represent data informativeness also prohibits its use in risk-conscious building performance management.

1.1.1.2 Deterministic parameter estimation

Instead of manually selecting calibration parameters and adjusting their values, automated model tuning relies on mathematical algorithms to infer those values in light of model agreement with observations. The most common automated model tuning approach in the literature formalizes calibration into a deterministic parameter estimation problem, uses standard statistical metrics as the objective function, and employs numerical algorithms to find the optimal, i.e. the parameter values that minimize this objective function. Within the scope of parameter inference techniques, this section focuses on the parameter space under exploration, the objective function being used, and the algorithm to

search for the optimal. Detailed reviews on the optimization techniques used in BPS applications can be found in the work of Machairas et al. (2014) and Nguyen et al. (2014).

In calibration of BPS models, Djuric et al. (2008) used a built-in sequential quadratic programming (SQP) algorithm to minimize the root mean square error (RMSE) between predictions and observations from a simple heat balance model. The entire set of parameters are calibrated since the model is relatively simple, and their upper and lower bounds were chosen based on on-site visit and authors' experience. Lavigne (2009) used the Marquardt-Levenberg's method to calibrate a model in DOE-2.1E by minimizing RMSE in the form of a quadratic function of important model parameters. This method calculates the associated Jacobian matrix used in the algorithm by running the physical model. A pre-calibration with a five-variable energetic model from ASHRAE Fundamental (ASHRAE 2011) is performed to identify the important subset of DOE-2.1E model parameters to be calibrated. Taheri et al. (2012) used a hybrid generalized pattern search with particle swarm optimization to calibrate a model in EnergyPlus (Crawley et al., 2000) by minimizing a weighted combination of R^2 and the coefficient of variation of root mean square error (CV-RMSE). The authors chose calibration parameters based on previous experiences, with sensitivity analysis as a suggested alternative. A similar approach was used by Tahmasebi and Mahdavi (2013). Ramos Ruiz et al. (2016) used genetic algorithm to minimize a combined metric of R^2 and the CV-RMSE. They applied sensitivity analysis based on both the "relative deviation" method and the Morris method (Campolongo et al., 2007) to select the calibration parameters, whose value ranges are equally-possible discrete values that are consistent with vendor specifications and building documentations.

Because of the high computation cost of common physical models, studies often use statistical surrogate models or meta-models to emulate the physical model within an optimization routine, which allows for exploration of a large parameter space and expedites the search of the optimal. O'Neill and Eisenhower (2013) used a gradient-based global optimizer to minimize the RMSE between observations and predictions from a support vector machine (SVM) meta-model of a physical model in EnergyPlus. They chose a small subset of calibration parameters based on a derivative-based sensitivity and arbitrary parameter value ranges. Robertson et al. (2015) applied a gradient-based simulated annealing optimization algorithm to calibrate a normal multivariate linear regression meta-model of a model in DOE-2.1E against synthetic utility data of a residential building. This study uses the CV-RMSE as the objective function, and calculates sensitivity coefficients by Monte Carlo simulation to select a subset of six calibration parameters. Yang and Becerik-Gerber (2015) proposed a comprehensive model calibration framework for simultaneous multi-level building energy simulation, which implicitly adopted normal multivariate linear regression emulator and quasi-multi-objective optimization with a weighted objective function and linear programming algorithm. This study performs classification of model parameters to differentiate estimable and adjustable parameter, and select important adjustable parameters for calibration using the Morris method.

Advances in numerical algorithms and supercomputing system technologies can greatly facilitate the above optimization-based calibration methods. A particular example is the "Autotune" calibration method (Chaudhary et al., 2016). This method employs an evolutionary algorithm to estimate the parameter values that minimize the standard normalized mean bias error (NMBE) and the CV-RMSE values. The evolutionary

operators used in this study include heuristic crossover, Gaussian mutation, tournament selection, and generational replacement. In searching for the true global optimum, the use of supercomputing systems and specially tailored search algorithms enables handling of hundreds of parameters and dozens-to-millions of measured data points. In addition, this study identified a set of 47–470 parameters most important for each building type using a priori sensitivity analysis based on existing simulations.

Within a conventional optimization framework, Reddy (2006) recognized the issue of uncertainty in calibration of detailed building simulation models, and conducted the ASHRAE RP-1051 research project (Reddy and Maor, 2006; Reddy et al., 2007a; Reddy et al., 2007b). They proposed a two-step approach to search within the parameter space for plausible solutions. The first step, a "bounded" coarse grid calibration involves a procedure that, similar to parametric and sensitivity analysis in manual model tuning, identifies both promising solutions of parameter values and parameters influential on the model discrepancy. The second step, a guided search calibration, looks for the final set of solutions through either manual or automated calibration methods. In particular, Sun and Reddy (2006) proposed an analytical parameter estimation method for the guided search, which used a gradient-based nonlinear optimization technique to minimize a weighted value of the CV-RMSE of both consumption and demand data as the objective function. This method identifies calibration parameters based on the normalized sensitivity coefficients determined by the perturbation method, a local sensitivity analysis approach. It also recognizes the mutual correlation among the calibration parameters to ensure they are mathematically identifiable. Different from the conventional approaches, the authors suggested using a set of plausible solution to evaluate energy conservation measures

(ECMs) to account for uncertainties in the model, which is a notable improvement over the CV-RMSE in generating meaningful probabilistic predictions. Nevertheless, this approach intrinsically assumes that all the plausible solutions are of equal probability regardless of either their agreement with the observations or their deviation from a "best guess" value. The lack of a solid statistical basis of this assumption compromises its effect in applications that were performed by Robertson et al. (2013) and Gestwick and Love (2014). Another common issue in model calibration is overfitting (Dietterich, 1995), where one obtains incorrect and often irrational parameter values in fitting observations as they subsume model bias. Attempts to address over-fitting in deterministic parameter estimation include the work of Carroll and Hitchcock (1993), which proposed to include a penalty term in the objective function. Taking the form of a sum of weighted square difference between a plausible solution and its corresponding default values intrinsically rejects those values deviating drastically from the default despite that they provide good agreement. A similar regularization approach, i.e. inclusion of penalty function to prevent over-fitting, was used by Lavigne (2009) and Nassiopoulos et al. (2014) where the Levenberg–Marquardt algorithm has a built-in tuning parameter to regularize the optimization.

Nevertheless, because of imperfect models and limited observations in real practice, this group of calibration methods based on deterministic optimization is prone to over-fitting issue. A few attempts to consider parameter uncertainties fail to further translate those uncertainties into model predictions to inform building performance management practice. Lack of proper recognition of the main sources of discrepancy and inefficient use of data impairs the confidence of those methods in real practice.

1.1.1.3 Bayesian inference

Bayesian statistics receives extensive studies in mathematical and statistical literature (Bal et al., 2013; Bayarri et al., 2007; Brynjarsdottir and O'Hagan, 2014; Conti and O'Hagan, 2010; Higdon et al., 2004; Kennedy and O'Hagan, 2001), and has also been applied to a wide range of practices in numerous scientific and engineering disciplines (Guillas et al., 2009; Wikle et al., 2001; Zhang and Arhonditsis, 2008), and particularly in building performance simulation applications (Booth et al., 2012, 2013; Chong and Lam, 2015; Heo et al., 2012; Li et al., 2016; Manfren et al., 2013; Tian et al, 2014). In general, Bayesian inference includes assigning prior distributions to influential model parameters, and conditioning on observations to obtain their posterior distributions using Bayes' theorem. One can obtain the full joint posterior distributions of calibration parameters through a full Bayesian analysis using Markov chain Monte Carlo (MCMC) (Metropolis, 1953) or the newer Hamiltonian Monte Carlo (HMC) (Duane et al., 1987). Otherwise, a less accurate approximation by maximum a posteriori estimation and normal approximation can be used to inform later analysis in the case of complex models and crude estimations. In cases where the likelihood function is implicit, e.g. it is difficult to calculate the probability of observations under certain parameter values, Approximate Bayesian Computation (Diggle and Gratton, 1984) can be used as an alternative to quickly approximate the rigorous Bayesian inference.

In practice, one can embed Bayesian statistics as a specific inverse parameter estimation method in a calibration framework. This method exploits the benefit of modeler's knowledge and experience by explicitly incorporating them into parameters' prior distributions. In addition, Bayes' theorem ensures the consistency in adjusting these beliefs based on the observations. As the conventional optimization-based calibration

methods and their extensions toward more rigorous uncertainty quantification can be regarded as a specific case of a general Bayesian calibration method, the Bayesian framework possesses coherence, generality and applicability in addressing model calibration problems.

1.1.2 Validation method

Validation of BPS models in general practice involves collection of observations and assessment of model predictions' agreement with these observations, both of which will be briefly reviewed and discussed in this section.

1.1.2.1 Data collection

Collection of data in building performance management practice often relies on practitioners' knowledge and experience. Manual calibration methods using extensive building information and data, albeit costly in time and labor, often leads to a model of high validity. This is probably because these methods often involve evidence-based forward parameter estimation at the level of building sub-systems. Well-designed procedures regarding the use of information and data throughout the calibration process also contribute to the success (Pan et al., 2007; Yoon et al., 2003). On the contrary, most studies on automated calibration in the literature use compliance with calibration standard, like ASHRAE Guideline 14-2002 (ASHRAE, 2002), as the sole validation criterion. Lack of sufficient information and data in validation may impair the model's credibility and obscure the benefit of automation. Therefore, a systematic way to integrate the strengths of both approaches deserves further investigations on the collection of information and data.

Regarding the type of information and data, Fabrizio and Monetti (2015) provided a taxonomy that includes a set of levels approximately in an ascending order of collection effort: utility bills, as-built data, site visit or inspection, detailed audit, short-term monitoring, and long-term monitoring. Raftery et al. (2011) proposed a similar information and data taxonomy with the focus on the hierarchy of reliability, which includes data-logged measurements, spots for short-term measurements, direct observation (site surveys), operator and personnel interviews, operation documents, commissioning documents, benchmark studies and best practice guides, standards, specifications and guidelines, and design stage information (e.g. the initial model). A related graphical summary can be found in the work of Coakley et al. (2011).

In addition to the source and cost of data, this dissertation classifies the monitoring data in building operation using temporal, spatial, and categorical scales. The temporal scale concerns the coverage and resolution of monitoring data regarding its temporal variability, mostly due to weather and usage scenarios. The most common type of data in building performance management is the monthly utility bill, as it is readily available and reliable in most cases. The use of hourly or sub-hourly data, such as those from smart meters, building management systems (BMS), in-situ monitoring, etc. receives more and more attention recently (Chaudhary et al., 2016; Djuric et al., 2008; Heo and Zavala, 2012; Liu and Liu, 2011; Nassiopoulos et al., 2014; Srivastav et al., 2013; Yang and Becerik-Gerber, 2015). This type of data is usually more informative than monthly data in model validation because of the embedded dynamic characteristics. However, the length of hourly or sub-hourly data used in common analysis is typically limited to a few weeks for spot monitoring, or only aggregated at the whole-building level for smart meter data, as it is

difficult for common methods to handle large amounts of data. This may affect the data's coverage of variations of weather conditions, and expose the model to extrapolation risks. In addition, hourly or sub-hourly data contains relatively large variations because of varying and often unknown building usage, and its strong temporal correlation makes it difficult to filter out those external variations in identifying the performance of building fabric and energy supply systems. These drawbacks may limit its usability in calibrating BPS models in real practice.

The spatial scale deals with the coverage and resolution of monitoring data with respect to its spatial variability, i.e. building typology and room functions. Because of limited implementation of sub-metering especially in existing buildings, most studies in BPS applications use consumption data at the whole building level, and apply a single-zone modelling assumption to rooms adjacent to each other and having similar functions accordingly. This ignores the variability in individual rooms, and may lead to incorrect observations with the presence of large variability.

Finally, the categorical scale concerns the type of monitored state variables, and is often directly related with the output of interest in building performance management. The most common type of output is power or energy use of electricity, gas, and/or other fuels. In contrast, room air temperature is less commonly monitored, as it is typically maintained by the heating, ventilation, and air conditioning (HVAC) system according to the thermostat setting, and therefore is often relatively constant and less informative. This may cause the room air temperature to be used more often as a model input rather than an output in real practice (Mustafaraj et al., 2014; Royapoor and Roskilly, 2015). However, if the indoor condition is less well maintained because of certain control logic, unexpected

building usage, equipment deterioration, system malfunction, etc., monitoring of the temperature (and possibly humidity as well) will also become necessary. ASHRAE Guideline 14-2002 pointed out this issue in Section D6 as well. Roberti et al. (2015) provided a case study that calibrates a model of a historical building to hourly indoor air and surface temperature. Ramos Ruiz et al. (2016) performed a calibration of a building envelope by comparing with interior temperature measurements. Regarding energy supply systems specifically, monitoring data of system state variables, like water and air temperature and flow rate, are usually more informative than consumption data as they reduce the scope of system of interest and block the external noises. This is an important reason that most procedure-based manual calibration can often render valid models.

In summary, model validation in real practice often builds upon its agreement with whole building monthly consumption data, which is often too aggregated and incomplete to reveal detailed dynamic characteristics. Monitoring data with very high temporal and spatial resolutions and/or belonging to other categories, on the contrary, could be either too noisy under uncontrolled and unknown weather and usage conditions, or less irrelevant from the output of interest. Furthermore, understanding of the mode validity under uncertainty from a risk-conscious decision making perspective is barely addressed in the literature. Therefore, a systematic method to assess data informativeness under uncertainty is worth further study for model validation in building performance management.

1.1.2.2 Accuracy metrics

In common practice, a BPS model is deemed calibrated if its prediction agreement with observations, i.e. goodness-of-fit, reaches a certain threshold. This agreement is often quantified by standard statistical metrics such as the NMBE and the CV-RMSE:

$$NMBE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{(n-p)\bar{y}} \tag{1}$$

$$CVRMSE = \frac{1}{\bar{y}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}} \tag{2}$$

where y_i and $\hat{y}_i$, $i = 1,2,\dots n$ are the observations and corresponding (point) estimates respectively, p is the lost degrees of freedom in the regression context and often takes the value of 1 in calibrating BPS models. This validation approach forms the calibration criteria established in ASHRAE Guideline 14-2002 (ASHRAE, 2002), the International Performance Measurement and Verification Protocol (IPMVP, 2002), and the Federal Energy Management Program (Webster et al., 2015) for observations at different time scales. The statistical basis of ASHRAE Guideline 14-2002 in particular, developed by Reddy and Claridge (2000), applies to quantification of uncertainty of energy savings in measurement and verification (M&V) analysis, where a normal linear regression model is created/calibrated to predict business-as-usual outcome in post-retrofit period. Under the assumptions of normal linear regression, one can translate these statistical metrics into uncertainty of energy savings and generate a crude yet rigorous estimation for saving verifications.

However, these calibration criteria based on goodness-of-fit of deterministic predictions are not suitable to evaluate BPS models with explicitly quantified uncertainties,

whose probabilistic predictions are more informative in risk-conscious BPS applications. ASHARE Guideline 14 (2002) states that, “There is still no broad consensus as to how to determine uncertainty or risk levels based on a calibrated simulation approach. Hence this annex has addressed this issue at a rather superficial level”. Reddy (2006) also recognized that “the methodology (to address uncertainty in ASHRAE Guideline 14-2002) applies to regression models identified from baseline monitored data and, hence, relates to black-box and grey-box approaches. It cannot be applied as such to the calibrated simulation model approach...” To be more specific, the connection between the prediction uncertainty of a model and its accuracy under these goodness-of-fit metrics is valid for normal linear regression models rather than BPS models because of their disparate model assumptions. The validity of using a t-distribution to translate the CV-RMSE into prediction uncertainties does not apply to an often under-determined BPS model with complex parameter correlations and limited observations, which makes the CV-RMSE based uncertainty ranges invalid and necessitates explicit uncertainty quantification in model calibration for risk-conscious BPS applications. Similar observations and a more technical explanation can be found in the work of Heo (2011).

Hence, the current standard calibration criteria cannot evaluate probabilistic predictions, nor can they inform risk-conscious decision making in BPS applications. Therefore, more informative and practical metrics to assess model accuracy and validity and corresponding data informativeness are needed in calibration of BPS models.

1.2 Research questions and hypotheses

The extensive review of research concerning calibration and validation of BPS models reveals two major research gaps. First, contemporary model calibration practice calls for an effective and robust method that can systematically incorporate a variety of information and data, handle modelling and prediction uncertainties, and maintain consistent model accuracy. Second, current approaches to collecting information and data in real practice largely depend on individual experience or common practice; further study is needed to understand the value of information and data, i.e. assess data informativeness, such as to support specific decision-making processes in choosing data monitoring strategies. This would avoid missed opportunities or wasted resources, and enhance practitioners' confidence in embracing a new generation of methods and tools in BPS applications.

Therefore, this dissertation targets the following two research questions:

(1) How to make optimal use of available information and data to calibrate a building performance simulation model under uncertainty?
(2) How to quantify the informativeness of information and data and validate a building performance simulation model for risk-conscious building performance simulation applications?

Accordingly, the main research hypothesis of this dissertation is: the proposed framework effectively addresses data informativeness for risk-conscious building performance simulation application. More specifically, it includes two hypotheses:

I. The proposed calibration methods make improved use of data in constraining uncertainty and improving prediction.

II. The proposed explicit risk assessment improves the representation of model validity and data informativeness under uncertainty.

1.3 Dissertation structure

The rest of this dissertation will be structured as follows. CHAPTER 2 provides an overview of the proposed framework with respect to its theoretical benefits. CHAPTER 3 and CHAPTER 4 each introduce a case study where the proposed framework is applied to demonstrate its practical effectiveness. The first is a design of an empirical validation experiment in the context of risk-conscious validation methodology. Forward uncertainty propagation using detailed measurements will be applied to constrain the associated uncertainties and reduce the risk of mis-validation. The second is a hypothetical intervention analysis of an existing building with hydronic heating. Inverse modelling using Bayesian inference will be employed to calibrate a dynamic simulation model with observation data varying in temporal and categorical scales, and assess the impact of these data on a risk assessment of decisions related to intervention outcomes. A summary of research conclusions and recommendations for future work is presented in CHAPTER 5.

CHAPTER 2. FRAMEWORK OVERVIEW

This chapter forms the theoretical foundation of the proposed framework and explains its effectiveness toward the specific question under study. It starts by identifying sources of model discrepancy, i.e. the gap between predictions and observations. Uncertainty quantification is then introduced as a method to reduce model discrepancy and construct probabilistic predictions. After that, interpretations of model calibration and validation with respect to probabilistic predictions are provided, under which context the effectiveness of the proposed calibration and validation methods will be explained in the end.

2.1 Source of discrepancy

It is well known that, "all models are wrong; some are useful" (Box, 1976). The difference between measured system outcome, i.e. observations, and corresponding model output, i.e. predictions, comes from two main sources. The first source is *sampling variability*: every observation is solely a *single instance* from the underlying process and therefore an incomplete observation of the reality. The second source is *imperfect model*: any model is an idealization of the real physical world with some level of abstraction, so it is incapable of perfectly characterizing the true process. While these two sources may be confounded in nature, a clear albeit arbitrary definition to distinguish them in practice would facilitate investigations and discussions. For this purpose, this dissertation adopts the concept of "ideal forecast" from measure theory. One can refer to the work of Wang (2016) for formal definitions and detailed explanations; only a brief interpretation is provided here: an ideal probabilistic forecast relative to an information set makes the best

possible use of that information, so one can regard an ideal forecast as equivalent to the true process under observation, enclosed in the information set. In this sense, as long as one can deem a probabilistic forecast ideal according to certain measures, the observed difference is solely the result of sampling variability. Therefore, this dissertation defines *model discrepancy* as *the deviation of a certain probabilistic forecast from an ideal one for a specific information set*; this definition only concerns imperfect model with respect to predictions.

As commonly known, "the portion of the world captured by the model is an arbitrary enclosure of an otherwise open, interconnected system" (Rosen, 1991). This dissertation attributes this imperfection, or equivalently error, to four main sources that go through the model prediction process. First, regarding a model as a collection of variables, spatially and temporally related to each other through mathematical formulas governed by physical principles, the first type of error, *model parameter error*, comes from erroneous representation of empirical quantities that appear as parameters in the modelling. The second type of error, *model form error*, comes from imperfect idealization and abstraction of physical mechanisms in the form of the mathematical formulas, i.e. the model functional form defined by Morgan et al. (2009). Similar to "model bias" or "model inadequacy", this type of error refers to the systematic error that persists even when the parameter values are correct. Furthermore, *computation error* appears in discretizing over spatial and temporal domains, solving algebraic equations, and executing numerical realizations. Analytical solutions of models in the form of ordinary or partial differential equations are free from this type of error, but only a very small set of models in real practice have explicit analytical solutions. Finally, inevitable *observation error* exists in measuring the actual outcome.

Figure 2.1 shows the sources of model discrepancy along with the prediction generating process.

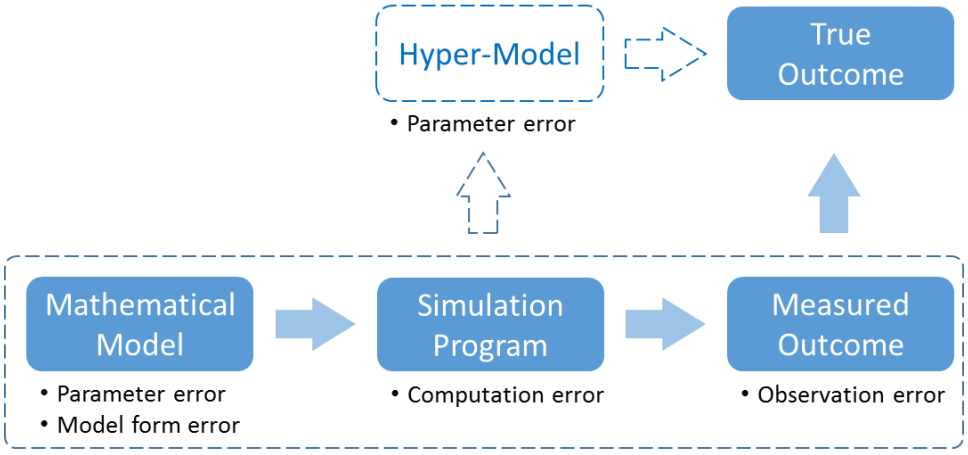


Figure 2.1 Sources of model discrepancy

This dissertation denotes by *hyper-model* a model that is supposed to generate an ideal probabilistic forecast, mostly by addressing all the four types of errors. A hyper-model is therefore assumed to be only prone to parameter error, since the model has explicitly parameterized and hence subsumed all the other types of error. Consequently, the probabilistic prediction of a hyper-model would be ideal once the parameter error is eliminated. This assumption helps avoid an otherwise endless process and, in model calibration, keeps the reduction of model discrepancy tractable. Figure 2.1 also depicts the relationship between a hyper-model and other previously mentioned entities.

2.2 Uncertainty quantification

Reducing model discrepancy requires identifying and minimizing the errors. A typical way is to represent the errors in the form of model uncertainties and quantify these uncertainties. Extensive research in identifying and classifying the sources of uncertainty

can be found in the literature (Kennedy and O'Hagan, 2001; Yao et al., 2011) and also in building energy modelling community (Hopfe, 2009; Rezaee, 2016; Sun, 2014; Wang, 2016). Considering uncertainty as another arbitrary idealization and abstraction of imperfect knowledge, this dissertation classifies sources of uncertainty from the modelling perspective as well; these sources consist of model parameter uncertainty, model form uncertainty, computation uncertainty, and observation uncertainty. While being mostly consistent with other classifications in the literature, a noticeable distinction in this dissertation is that, parametric variability (Kennedy and O'Hagan, 2001) or model input uncertainty (Yao et al., 2011), is considered as an external uncertainty that does not contribute to the model discrepancy. In other words, the capability to describe correctly the external conditions is not within the scope of a model: using the model to answer the "what if" question does not need to concern the correctness of the presumption "if".

One of the most popular ways to address model uncertainty employs probability theory (Yao et al., 2011), which represents uncertainty as random variables or stochastic processes, and specifies the according (joint) probability distribution using hyper-parameters, i.e. parameters that describe the probability distribution of model parameters. One can then construct probabilistic model predictions by propagating those uncertainties into model outputs through either analytical calculation or numerical sampling. From this perspective, this dissertation regards uncertainty quantification as *the process of characterizing model errors in the physical process by specifying and propagating model uncertainties*. This process includes both uncertainty propagation, i.e. the quantification of uncertainties in system outputs propagated from uncertain inputs, and inverse modelling, i.e. the estimation of uncertain inputs from observations of system outputs. Using the

concept of a hyper-model, quantification of all the sources of uncertainties associated with the original model is equivalent to creating a probabilistic hyper-model and generating probabilistic predictions. This intrinsically reduces model discrepancy at the same time.

2.3 Model calibration

An important step in building performance management especially in complex commercial buildings is to create a model to support intervention analysis under certain virtual experiment settings. This model belongs to the category of prognostic models according to Saltelli et al. (2008). Given an as-designed or as-built model, ASHRAE Guideline 14-2002 defines calibration as "the process of comparing the output or results of a measurement or model with that of some standard, determining the deviation and relevant uncertainty and adjusting the measuring device or model accordingly". Kennedy and O'Hagan (2001) provided a similar definition, where "the process of fitting the model to the observed data by adjusting the parameters is known as calibration". Another definition specific to simulation models is from Reddy (2006): "calibrated simulation is the process of using an existing building simulation computer program and 'tuning' or calibrating the various inputs to the program so that observed energy use agrees closely with that predicted by the simulation program".

To generically, albeit pragmatically, define BPS model calibration in the context of an ideal forecast, the link between the model and an information set needs to be established. This dissertation denotes by data the measured state variables, and denotes by information the rest of the available knowledge specific to the building under study, including general building information, as-designed or as-built drawings, manufacturer specifications, and

as-operational information obtained from building auditing. Meta-information refers to generic knowledge such as common practice, default values in handbooks and guidelines, modeler's experience, and the fundamental physical principles. Table 2.1 provides a summary of the classification. From this perspective, a model before calibration is a representation of an original information set including both information and meta-information; a calibrated model, on the contrary, is a representation of an updated information set enhanced by additional information and data from building auditing, utility bills, monitoring, etc.

Table 2.1 Information set classification

Information	Data	Meta-information
• General building information • As-designed/built drawings • Commissioning documentation • Operation and Maintenance (O&M) manuals	• Utility bills • Building management system (BMS) censor data • Spot measurement • Short/long term monitoring • Intrusive testing	• Physical principles • Guides and standards • Common practice • Modeler's experience

Within this framework, this dissertation defines model calibration as *the process to construct an ideal probabilistic forecast relative to an information set through model uncertainty quantification*, which is the same process as creating a hyper-model. This definition considers both forward and inverse uncertainty quantification, i.e. uncertainty propagation and inverse modelling respectively. This definition helps distinguish between calibration methods and calibration methodology. The former primarily concerns the techniques for explicit parameter inference and model identification, whereas the latter also

considers how particularly one should enhance the information set to support analysis; this enhancement includes what and how to collect the information and data, as well as to use which type of model, as part of the meta-information.

As a common model calibration method, Bayesian statistics has been widely used in many scientific and engineering disciplines. The fundamental principle of Bayesian statistics is Bayes' theorem:

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} = \frac{P(y|\theta)P(\theta)}{\int P(y|\theta)P(\theta)d\theta} \tag{3}$$

in which $P(\theta|y)$ is the probability of model parameter θ conditioned on observation y; $P(\theta)$ is the probability of θ without knowing y; $P(y|\theta)$ is the conditional probability of y given θ, i.e. the likelihood; $P(y)$ is the probability of y, i.e. the marginal probability, which is the expectation of probability of y over the distribution of θ. Bayes' theorem updates the estimate $P(\theta)$ to $P(\theta|y)$ by considering the contribution of information in y. Therefore, $P(\theta)$ is called the prior probability and usually presents general knowledge, and $P(\theta|y)$ is called the posterior probability and represents updated knowledge based on observations. This estimation technique is Bayesian statistics, and the Bayesian procedures in obtaining the posterior distributions for a given model and observations fall within the calibration method domain as Bayesian inference.

The theoretical soundness of Bayesian inference in model calibration naturally stands out within the context of previous definitions. First, Bayesian statistics is in principle probabilistic: it treats all unknowns as random variables, quantifies all the aspects of uncertainty via probability, and makes inferences using probability statements. This allows

a straightforward implementation of uncertainty quantification in risk-conscious decision-making. Second, one updates the prior distributions based on the observations using Bayes' theorem, which is intrinsically an inverse approach in the sense that the observations always inform the posterior distribution. Third, the classic Bayesian calibration framework from Kennedy and O'Hagan (2001) creates a hyper-model that considers all the sources of uncertainties simultaneously. Finally, a more important feature of Bayesian inference is that its fundamental principle inherently satisfies the idealness requirement in a more general manner; the Bayes' theorem ensures that the Bayesian inference, as a probabilistic forecast, is consistent with both the prior distributions and the observations. As the prior often comes from meta-information as well, one can deem the Bayesian inference to be ideal relative to the grand information set that consists of information, data, and meta-information. This further ensures that Bayesian inference makes improved use of data and therefore can effectively represent data informativeness in BPS applications.

In addition, the effectiveness of Bayesian inference also comes from its generality. As mentioned previously, conventional automated calibration methods use optimization algorithms to minimize one or several statistical metrics. Common choices of such include the NMBE and the CV-RMSE. It is easy to show that the CV-RMSE reduces to the NMBE if treating the sum of the entire data as a single point. This observation makes the CV-RMSE, equivalently the squared error, the general objective function in conventional calibration methods. A weighted combination of the NMBE and the CV-RMSE, such as the one proposed by Reddy et al. (2007a), is therefore equivalent to a weighted sum of the squared difference between predictions and observations at different scales or of different outcomes.

Given the above observation, statistical literature has shown that, "(in normal linear regression) under the standard non-informative prior distribution, the Bayesian estimates and standard errors of regression coefficients coincide with the classical (least-square) results." (Gelman et al., 2013). This comes from the fact that, 1) using non-informative priors ensures that the prior probability is constant over the plausible range of parameter values; 2) assuming a Gaussian error makes the likelihood function proportional to the squared difference between predictions and observations. Therefore, maximizing the posterior probability, i.e. the product of prior probability and likelihood, is equivalent to minimizing the squared difference. In the meantime, since the additional penalty function also usually takes the form of a weighted squared difference between the proposed parameter value and its default (Carroll and Hitchcock, 1993), a Bayesian counterpart can be created by constructing an equivalent Gaussian prior distribution with a mean being the default value and a variance being proportional to the weight. In this sense, if the solution is unique such as in a convex optimization problem like normal linear regression, one can expect the posterior mode in Bayesian inference to coincide with the least square estimate in deterministic parameter estimation.

Therefore, Bayesian inference generalizes deterministic parameter estimation techniques, and guarantees the optimal result in convex problems when a Gaussian error is assumed. In addition, the use of likelihood function allows a variety of forms of error beyond Gaussian error to be modelled in model calibration, which makes Bayesian inference a more flexible inverse modelling method. As in non-convex problems, the actual results largely depend on the problem set-up and the specific Monte Carlo (MC) algorithm being used, specific case studies are warranted to evaluate its effectiveness. In the

meantime, while Bayesian inference ensures the idealness relative to the priors and observations, those priors is subject to modelers' bias and could be incorrect for the actual conditions. To determine if a model is truly calibrated for its intended purpose falls inside the scope of model validation, which will be addressed in the following section.

2.4 Model validation

A formal definition of verification versus validation by Schlesinger (1979) states: verification is the process of comparing a computerized model with a conceptual model, while validation compares a computerized model with reality. This definition primarily agrees with the concept of empirical validation in the building performance simulation community (Judkoff and Neymark, 2006), and indeed forms the basis of common validation criteria like ASHRAE Guideline 14-2002. Particularly for computer simulation models, Sargent (2011) defined model validation to mean "substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model". In the context of an ideal forecast, this dissertation defines model validation as *the process to assess the idealness of a probabilistic forecast of a model relative to an information set for its intended application.* From a risk-conscious perspective, this definition emphasizes not only the forecast idealness that relates to model accuracy, but also the validity of a model to fulfill its intended purpose that may have to be addressed in a case-by-case manner with specifically defined risk measures.

Model validation in real practice involves determining what type of information and data should one collect, which leads one to the focus on the adequacy of the

information set. In other words, one can only expect a model to be valid for its intended application when the model generates ideal probabilistic forecasts relative to an adequate information set; the adequacy of the corresponding information set is thus a necessary, and probably also a sufficient condition of a valid model. Therefore, model validation also involves enhancing the information set until it achieves adequacy for a certain application, which in turn defines data informativeness as *the value of data toward the adequacy of an information set for an intended application*. If a model is ideal relative to an information set, the model's validity for an intended application solely depends on the adequacy of the information set, and thus also on the sufficiency of the informativeness of data. This is how data informativeness is related to model calibration and validation, and this relation forms the theoretical foundation of the proposed framework.

While model validity can only be addressed in a case-by-case manner, model accuracy can typically be evaluated using certain metrics. As the NMBE/CV-RMSE is not sufficient for evaluating probabilistic predictions with respect to idealness, an effective framework needs to include appropriate model accuracy metrics. Building upon the concept of an ideal forecast, a promising candidate is marginal and probabilistic calibration (Gneiting et al., 2007), where calibration refers to the statistical consistency between the distributional forecasts and the observations. This evaluation approach comes from the traditionally used probability integral transform (PIT) (Rosenblatt, 1952) which tests the agreement between a specified single distribution and a set of observations. Gneiting and Katzfuss (2014) generalized PIT to allow the testing of a set of distributions by proposing the definition of a prediction space, in which each pair of predictive cumulative distribution function (CDF) F and a corresponding observation y constitutes an element (F, y). Albeit

being necessary but not sufficient conditions for an ideal forecast, it is appropriate to treat marginal and probabilistic calibration as strong requirements in practice. Related formal definitions and interpretations for further understanding can be found in the work of Wang (2016).

While the above concept of calibration is a joint property of the predictions and observations, sharpness refers to the concentration of the predictive distributions and is a property of the predictions only. Within the context of model calibration, sharpness has a natural link to prediction uncertainty. This link further reinforces the expectation that enhancing the information set reduces uncertainty, as an ideal forecast relative to a larger information set is generally sharper. Therefore, it is equivalent to consider model calibration as to "maximize the sharpness of the predictive distributions subject to calibration" (Gneiting et al., 2007).

From this point of view, the potential of the continuous rank probability score (CRPS) as a scoring rule to quantitatively assess model accuracy stands out, as this score considers both calibration and sharpness at the same time. Measuring the agreement of predictive distributions and corresponding observations, this score is widely used in forecast verification (Gneiting and Raftery, 2007) and also in evaluating building performance predictions (Li et al., 2016; Sun, 2014). The CRPS is defined as:

$$CRPS(F, y) = -\int_{-\infty}^{\infty} \big(F(t) - \mathbb{1}(t - y)\big)^2 dt \tag{4}$$

where $\mathbb{1}$ is the Heaviside step function and denotes a step function along the real line that attains 1 if $t - y \geq 0$ and 0 otherwise. In practice its negative orientation is typically used,

say $CRPS^*(F,y) = -CRPS(F,y)$. This dissertation denotes by the CRPS its negative orientation thereafter for simplicity. An illustrative example can be found in Figure 2.2 (Hamill, 2007), where the value of the CRPS is related to the shaded area in the figure on the right.

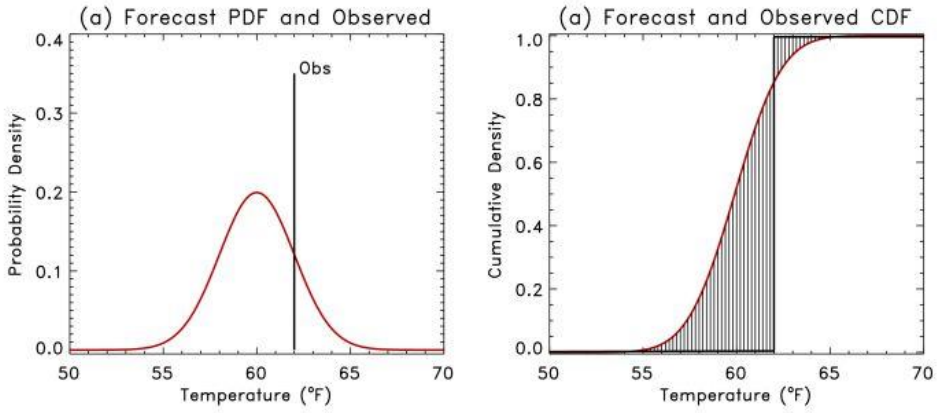


Figure 2.2 Illustration of the meaning of the CRPS

The CRPS is expressed in the same unit as the observed variable. A large value of the CRPS indicates a large discrepancy between the predictive distribution and the single observation. One can calculate the score of a probabilistic prediction from Monte Carlo simulation by:

$$CRPS(F,y) = E_F|Y - y| - \frac{1}{2}E_F|Y - Y'| \tag{5}$$

where F is the predictive distribution of random variable Y represented by the sample set, y is the single observation, E_F is the expectation over F, Y' is an independent random variable with identical distribution as Y, obtained by random permutations of the sample set F. Equation 5 shows that the CRPS equals the expected absolute error with respect to the single observation, minus half of the expected absolute error due to its own variability

as a measure of prediction reliability (Hersbach, 2000). Equation 5 also shows that the CRPS generalizes the absolute error, to which it reduces if F is a point forecast (Gneiting and Raftery, 2007). Figure 2.3 shows the CRPS of a normal distribution prediction of a true observation with different magnitudes of bias and standard deviation, which clearly shows that the CRPS increases when the probabilistic prediction is too far from observation, with the prediction bias, i.e. a constant offset between the mean of the normal distribution and the true observation, increases from 1 to 5. At the same time, the CRPS also penalize a prediction with too large dispersions, represented by the standard deviation of the normal distribution increasing also from 1 to 5. One can refer to Gneiting and Raftery (2007) for more details about the CRPS for further understanding.

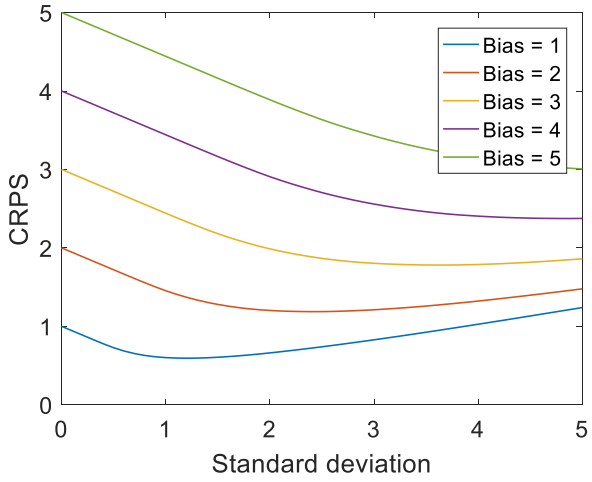


Figure 2.3 Illustration of the CRPS with normal distribution

For a group of observations and their corresponding probabilistic predictions, the mean CRPS, optionally normalized by the observation mean, can be used accordingly as in Equation 6 and 7. This dissertation will investigate the practical effectiveness of the CRPS in assessing model accuracy and quantifying the related data informativeness in risk-conscious BPS applications.

$$MCRPS(F,y) = \frac{1}{n}\sum_{i=1}^{n}\left(E_F|Y_i - y_i| - \frac{1}{2}E_F|Y_i - Y_i'|\right) \quad (6)$$

$$nMCRPS(F,y) = \frac{1}{\bar{y}}MCRPS(F,y) \quad (7)$$

2.5 Summary

This chapter lays out the theoretical foundation of the proposed framework from the perspective of probabilistic predictions, which ties together model calibration, model validation, and data informativeness via the concept of ideal forecasts. In summary, model calibration aims to achieve the idealness relative to an information set through uncertainty quantification, and model validation examines the presumably ideal model against its intended application through explicit risk assessment, which in turn quantifies the informativeness of data relative to which the model achieves idealness. The rest of this dissertation will provide two case studies to demonstrate this framework's practical effectiveness in terms of the forward and inverse uncertainty quantification methods, the CRPS as model accuracy metrics, and explicit risk assessment to represent data informativeness.

CHAPTER 3.

EMPIRICAL VALIDATION EXPERIMENT DESIGN

Uncertainty propagation quantifies the uncertainty of model inputs and propagates them into uncertainties of system outputs. In model calibration in real practice, uncertainty propagation is typically achieved by reducing the scope of system of interest and monitoring key state variables through extensive sub-metering, site visits, short-period monitoring, and in-situ tests. Although this approach can often render reliable estimates of model inputs, the cost of time and labor associated with reducing the scope of system of interest and monitoring state variables usually prohibits its application in large and complex buildings under uncontrolled usage conditions. In contrast, empirical validation experiments are usually performed in single-room test cells and under controlled or closely monitored weather and usage conditions. The extensive and high resolution monitoring of a variety of state variables in an empirical validation experiment also enables direct estimation of key model inputs and assessment of informative data. Hence, this dissertation chooses a series of future empirical validation experiments as the first case study to demonstrate and test the proposed framework, with a focus on the uncertainty propagation techniques, identification of informative data, and model validation.

3.1 Background

Empirical validation, as with analytical verification and inter-program comparison, has been widely recognized as a useful method to validate a simulation program (Judkoff and Neymark, 2006). However, the complexity and large cost of time and labor of

empirical validation lead to only a few related studies in the literature, mostly in single-room climate chamber or test cells experiments, including IEA Annex 21 (Lomas et al., 1997), IEA Task 22 (Palomo del Barrio and Guyon, 2003, 2004), and the PASSYS project (Clarke et al., 1993; Jensen, 1993). A noticeable exception that applies to realistic full-scale residential building can be found in IEA Annex 58 (Strachan et al., 2015; Strachan et al., 2016; Strachan et al., 2015). Although, because of uncertainties in the building characteristics, they ended up focusing on only one room in each of the two buildings.

While the literature sees the maturity of design of experiments and procedures in empirical validation, only a few studies have paid special attention to the handling of model uncertainties. Palomo et al. (1991) suggested using residual analysis to assess simulation accuracy, identify discrepancy causes, and improve model prediction. They proposed a variety of statistical metrics to quantify the discrepancy and inform model improvement. Palomo and Guyon (2003, 2004) proposed a two-step empirical model validation methodology that includes checking model validity and diagnosis. The first step relies on residual analysis and comparison between uncertainty intervals of model outputs and measurements. Based on a linear assumption, the second step uses local sensitivity analysis to identify parameters influential on model discrepancy, and uses optimization techniques to search for parameter values that minimize the discrepancy. Strachan et al. (2016) used the absolute difference and Pearson correlation coefficient to assess the magnitude of profile agreement of model prediction.

In addition, despite well-established information and data collection procedures, the use of data remains insufficient to guarantee an adequate empirical validation experiment under uncertainty. A noticeable exception appears in the work of Clarke et al.

(1993) in the PASSYS project, which involves the use of high quality and high resolution data of different categories to validate and calibrate a simulation model of two test cells in ESP-r. This study uses surface temperature and heat flux measurements to identify possible reasons for large residuals, and obtains model parameter estimates for a better agreement. Testing of this component-level inverse parameter estimation against room-level air temperature shows a good level of agreement. Nevertheless, further studies are needed to understand the validity of the overall experimental design in terms of its collection and use of data under uncertainty, such as to ensure its adequacy in validating BPS models for general building performance management purpose.

3.2 The proposed empirical validation methodology

A risk-conscious empirical validation methodology needs to be able to handle uncertainties associated with the experiment set up, such that one can appropriately attribute the discrepancy between model predictions and observations to parameter and model form errors, and accredit the model's accuracy accordingly. Conventional empirical validation is based on comparing this discrepancy with a "standard" threshold. Because of the associated uncertainty in both the model and the experiment, a poorly designed experiment with extremely large uncertainties in the real building characteristics may not be able to distinguish valid and invalid models, as the differences between these models are overwhelmed by the errors due to those uncertainties. Hence, identifying and reducing parameter uncertainties should become a primary concern in the experimental design and validation methodology of empirical validation studies.

Therefore, this case study proposes to construct an *internal* (i.e. used only by the experimenter) probabilistic model prediction that fully considers all sorts of uncertainty by uncertainty propagation, and then use its model discrepancy, presumably the "best guess", as the threshold instead to assess other models, i.e. *external* models. In addition, the discrepancy of this "best guess" itself becomes an indicator of the adequacy of the empirical validation experiment. If a large discrepancy between the internal model and the observations is present, one cannot expect this experiment to be able to detect and reject poor external models.

From this perspective, the proposed empirical validation methodology includes the following steps:

1) Given the available facility information and experiment setup, create a corresponding internal simulation model, and identify and quantify the uncertainty of all the model parameters based on empirical knowledge and monitoring data.
2) Generate a probabilistic prediction of the system output by non-intrusively propagating parameter uncertainty through experimental design, random sampling, and repeated simulations.
3) Assess the validity of the experiment by examining this prediction's agreement with observations based on probabilistic accuracy metrics. Large discrepancy indicates the inadequacy of the experiment, along with its information, setup, and measurements, to validate an external simulation model. Measures to constrain associated uncertainty and reduce the discrepancy, for example an

experiment repeat with more detailed measurements, should then be taken to refine the experiment.

4) Once validated, the level of agreement will serve to assess and validate external simulation models.

3.3 Experiment description and simulation models

Lawrence Berkeley National Laboratory (LBNL) is going to perform a series of experiments to benchmark current building performance simulation of a single-zone conventional mixing system. A series of future experiments, the idealized low mass, south-facing window tests of air handling unit heating/cooling (AHU heating/cooling, I-600-H/C) that resemble BESTEST (Judkoff et al., 2010) #600 tests constitute the first case study. The 20×30×12ft. test cell X-3B is used for the targeted experiments. Its south exterior wall and window are fully configurable. The rest of the envelope, including the partition wall to test cell X-3A on the west, a temporary ceiling and north partition wall added to remove construction complexities, and the east exterior wall are equipped with R-80+ insulation. The test cell has a radiant slab with embedded hot water tubes, whose temperature will be maintained to be equal to the anticipated room temperature to reduce heat transfer. Additional insulation board will also be added on the top of the radiant slab during the experiment. Air injection with constant flow rate will be applied to the test cell to maintain a positive pressure and eliminate infiltration.

I-600-H/C tests will use the AHU water coils to maintain the cell air temperature at a pre-determined setpoint, either constant (CT) or with a setback (ST) from 6:00 pm to 8:00 am on the next day. Each set of tests also include a free-floating (FF) test where AHU

conditioning is turned off. As the up-to-date experiment plan is not available yet, this case study assumes the setpoint to be 30°C with an optional 10°C setback in the heating tests, and 20°C with an optional 10°C setback in the cooling tests. Each test is assumed to last six days with the first three days as warm-up, and the entire set of tests will be performed on 07/05-07/22 in the summer and 12/05-12/22 in the winter of a typical meteorology year (TMY). Monitoring of the experiment conditions and outcomes includes:

- Cell: air and operative temperature at multiple locations.
- Envelope: interior/exterior surface temperature and heat flux at multiple locations.
- Window: incident and incoming vertical insolation
- Exfiltration: constant injection flow rate and air temperature
- AHU: air/water flow rate and inlet/outlet temperature
- Meteorological conditions: on-site dry bulb and dew point temperature, global and diffusive irradiance, wind speed and direction, global infrared radiation, and ground reflected insolation.

As the actual experiment measurements are not yet available, a “true” model in EnergyPlus 8.7.0 is used to serve as a surrogate of reality and generate synthetic measurements (Figure 3.1). This model is created based on facility drawings and experiment descriptions, and has perturbations relative to specifications on parameter values of a variety of model inputs. All the adjacent rooms are neglected, and the boundary conditions are specified by predetermined surface temperature profiles plus random noises on the other side of the partition walls and the temporary ceiling. These random noise are based on the sensor error from manufacturer specifications, and therefore are assumed

Gaussian. Complexity of construction in terms of corrugation, studs, box-in columns are neglected. The radiant slab is modelled in the same way as the partition walls, i.e. by predetermined boundary surface temperature profiles with random noise. Modelling of the ground, the structural slab and the tube-embedded radiant slab is therefore neglected; only the insulation panel, below which 9 surface temperature sensors are to be installed, is modelled as the sole floor construction. Air injection is modelled as zone ventilation with constant and known flow rate, and is assumed to have the same status with outdoor air because of lack of information.

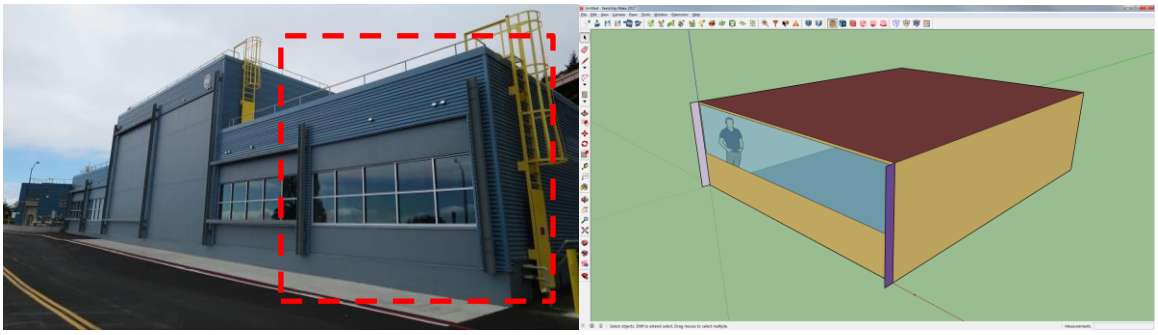


Figure 3.1 FLEXLAB test cell X-3B and its model in Energyplus

The "true" model also considers potential room air stratification by using the EnergyPlus object *RoomAir:TemperaturePattern:ConstantGradient*, which defines a constant vertical room air gradient along with offsets between mean room air temperature and thermostat, return and exhaust air temperature. The actual AHU system is modelled as an idealized system in EnergyPlus to remove system uncertainty.

A baseline internal model is constructed similarly, which differs from the "true" model by using the specifications directly for a variety of model inputs. Surface boundary temperatures are also replaced with synthetic measurement data. In addition, the "measured" cell mean air temperature is used as the thermostat setpoint in the heating and

cooling tests, such that one would expect the computed system heating and cooling power to match the synthetic observations from the "true" model. The following uncertainty quantification based on generic knowledge is then performed on this model to construct the baseline probabilistic predictions.

3.4 Baseline uncertainty quantification and sensitivity analysis

Quantification of the system output uncertainty of the baseline internal model builds upon a generic uncertainty quantification repository (Sun, 2014; Wang, 2016) based on empirical knowledge, existing literature, and previous experience. At this stage, only synthetic measurements related to thermostat setpoint and boundary conditions are used and their uncertainties are quantified and propagated. This is to represent the baseline case where more detailed measurements are not available, so the uncertainties associated with system outputs reflect the reasonably possible outcome from an external simulation model given the same amount of information, and only an exceeding discrepancy, probably due to large model errors or modelling mistakes, indicates that the model is invalid.

3.4.1 Material properties

Uncertainty information of envelope material properties comes primarily from the work of Macdonald (2002). All the other parameter uncertainties are assumed based on previous experience and knowledge of similar properties of other materials. Bounded normal distribution are assumed for each property parameter, whose standard deviations are summarized in Table 3.1.

Table 3.1 Uncertainty of material properties

Model parameter	Standard deviation	Unit
Opaque material		
Conductivity	5%	W/m/K
Density	1%	kg/m^3
Specific heat	12.25%	J/kg/K
Thermal absorptance	2%	-
Solar absorptance	7%	-
Visible absorptance	7%	-
Air gap		
Thermal resistance	5%	$m^2/K/W$
Glazing		
Solar transmittance	1%	-
Front side solar reflectance	1%	-
Back side solar reflectance	1%	-
Visible transmittance	1%	-
Front side visible reflectance	1%	-
Back side visible reflectance	1%	-
Infrared transmittance	1%	-
Front side infrared emissivity	1%	-
Back side infrared emissivity	1%	-
Conductivity	5%	W/m/K
Dirt correction factor	10%	-

3.4.2 Convective heat transfer coefficient

Quantification of uncertainties in convective heat transfer coefficients of interior and exterior surfaces of the building envelope employs the approach proposed by Sun (2014). This approach is based on the DOE-2 convective heat transfer model, and quantifies the uncertainty associated with model coefficients concerning natural and forced convection, i.e. stack and wind effects. Bivariate joint distributions are derived using meta-

analysis for vertical wall, floor, and ceiling surfaces individually. These uncertainties are added by perturbing inputs of EnergyPlus objects *Curve:Linear/Exponent* and applying these curves to the corresponding convective heat transfer coefficients through objects *SurfaceConvectionAlgorithm:Inside/Outside:UserCurve*. Variability of those coefficients among individual interior wall surfaces is neglected for simplicity, so they share the same convective heat transfer model coefficients in the same simulation instance. More details about the uncertainty quantification method can be found in the work of Sun (2014).

3.4.3 Sensor error

Table 3.2 summarizes the sensor errors based on manufacture specifications, which apply to all the sensors without calibration. In case calibration of sensors is to be performed before the experiment, the reference sensor reading is assumed to have no bias and rated error bounds, whereas the to-be-calibrated sensor reading has an individual constant bias and a random error. Let v^t denotes the true physical value to be measured at time step t, $t = 1,2, \dots k$; δ_{c^i} denotes the bias of the ith to-be-calibrated sensor, $i = 1,2, \dots n$; ε_r and ε_{c^i} denote the random error of reference and to-be-calibrated sensor reading respectively, then for the reference sensor,

$$v_r^{t\prime} = v^t + \varepsilon_r, \varepsilon_r \sim \mathcal{N}(0, \sigma_M^2) \tag{8}$$

In which σ_M^2 can be estimated from manufacturer specifications, as shown in Table 3.2, by assuming the range is equal to three times of the standard deviation. For the to-be-calibrated sensor,

$$v_{c^i}^{t\prime} = v^t + \varepsilon_{c^i}, \varepsilon_{c^i} \sim \mathcal{N}\left(\delta_{c^i}, \sigma_{c^i}^2\right) \tag{9}$$

Therefore, their difference at each time step has:

$$\Delta^t = v_{c^i}^{t\prime} - v_r^{t\prime} = \varepsilon_{c^i} - \varepsilon_r \sim \mathcal{N}\left(\delta_{c^i}, \sigma_{c^i}^2 + \sigma_M^2\right) \tag{10}$$

which gives the estimates of δ_{c^i} and $\sigma_{c^i}^2$ as:

$$\hat{\delta}_{c^i} = \bar{\Delta}^t, \hat{\sigma}_{c^i}^2 = \frac{1}{k-1}\sum_{t=1}^{k}(\Delta^t - \bar{\Delta}^t)^2 - \sigma_M^2 \tag{11}$$

Table 3.2 Sensor error

Sensor type	Manufacturer error	Unit
Test cell conditions		
Surface temperature	±0.05	°C
Air temperature	±0.05	°C
Water temperature	±0.03	°C
Water flow rate	±0.41%	m^3/s
Meteorological conditions		
Dry bulb temperature	±0.05	°C
Dew point temperature	±0.2	°C
Global irradiance	±5%	W/m^2
Diffuse irradiance	±5%	W/m^2
Wind speed	±0.1	m/s
Wind direction	±1	°

3.4.4 Meteorological conditions

The sensor error is considered as the only source of uncertainty in meteorological conditions because of the use of on-site weather station and local measurements. Uncertainty of ground reflectance, on the other hand, is assumed from 0.1 to 0.4 based on experience at this stage. Table 3.2 also summarizes the associated uncertainties. This case study uses TMY weather file from the nearby Oakland International Airport as the synthetic measurements, and draws sample from sensor error bounds to generate weather file instances.

3.4.5 Temperature as boundary conditions

Surfaces with adjacent boundary conditions in the experiment use synthetic temperature measurements as a time-varying boundary condition. It is believed that the uncertainty of an individual sensor comes only from time-independent sensor errors. First, each surface temperature sensor reading $T_{c^i}^{t\prime}$ follows:

$$T_{c^i}^{t\prime} = T^t + \delta_{c^i} + \varepsilon_{c^i} \sim \mathcal{N}\left(T^t + \delta_{c^i}, \sigma_{c^i}^2 + \sigma_M^2\right) \tag{12}$$

where T^t is the true temperature. Acknowledging the non-uniform temperature distribution on the surface, the area-weighted mean value follows:

$$\hat{T}^t = \frac{1}{A}\sum_{i=1}^{n} A_i \left(T_{c^i}^{t\prime} - \hat{\delta}_{c^i}\right), Var\left(\hat{T}^t\right) = \frac{1}{A^2}\sum_{i=1}^{n} A_i^2\left(\sigma_{c^i}^2 + \sigma_M^2\right) \tag{13}$$

where A_i is the representative area of each sensor reading, and unknown parameters are to be replaced by estimates from manufacture specifications and sensor calibration. Uncertainty associated with cell temperature measurements can be estimated in the same way as in Equation 13, except using volume-weighted mean instead. Similarly, in the case

of comparing predictions and observations aggregated on the time resolution, for example the hourly mean value, to aggregate sub-hourly readings within each hour t_s follows:

$$\hat{T}^{t_s} = \frac{1}{s}\sum_{j=1}^{s} \hat{T}^{j}, Var\left(\hat{T}^{t_s}\right) = \frac{1}{s^2}\sum_{j=1}^{s} Var\left(\hat{T}^{j}\right) \quad (14)$$

Equation 13 and 14 clearly show that aggregation over space and time constraints the uncertainty/variance. Surface boundary conditions are determined by averaging over 9 synthetic temperature sensor readings with equal representative area on each surface; the random error variance is therefore one third of a single sensor. The cell air temperature has 27 sensors per the experiment plan, and the uncertainty of the mean is computed accordingly. A Monte Carlo sample of four hundred random sequences of each surface and the cell mean temperature is drawn by independent sampling at each time step using Latin Hypercube design (LHD), and is then imported into the model by object *Schedule:File* and *SurfaceProperty:OtherSideCoefficients*. In uncertainty propagation, a random sample is drawn from a $(0,1)$ uniform distribution to determine the sequence to be used as the temperature boundary condition in each simulation instance.

3.4.6 Room air stratification

Based on the generic uncertainty quantification repository, a ±1°C uncertainty is added to the offsets of thermostat, return and exhaust air temperatures, whereas the constant room vertical temperature gradient is assumed to range from 0.5 to 1.875 °C/m.

3.4.7 Output measurements

Uncertainty quantification of cell mean air temperature measurements in free-floating tests follows the same way as above. Uncertainty in heating and cooling power measurement comes from sensor errors of air and water temperature and flow rate. Table 3.3 summarizes their respective computed error bounds.

Table 3.3 Uncertainty of system outputs

Measured outcome	Error	Unit
Room temperature	±0.001	°C
Heating/cooling power	±0.5%	W

3.4.8 Baseline model analysis result

The baseline uncertainty propagation uses a LHD sample of 2000 points on 148 parameters, including real uncertainty parameters and those samplers that are used to select sample sequences. Results of I600-H/C-ST tests against synthetic output measurements are shown in Figure 3.2 as an example, and the full results are shown in Figure A.1 in APPENDIX A. The results show overall good agreement of the mean prediction (black line) with the observation (blue line), but have considerably wide uncertainties (grey band), suggesting a potential large error an external model would possess even with reasonable estimates of parameter values according to the available information and data. This in turn reflects the invalidity of the baseline internal model as well as the inadequacy of the current information set in validating external models. A further calibration of the internal model with more detailed measurement becomes necessary.

It is worth noting that the large deviation occurring in the second and third day of the two heating tests is probably because of the sensitivity of heating power to setpoint

settings in a mild west coast climate. In this case, all the second-order, difficult-to-model effects become relatively more important, and their uncertainties lead to the large uncertainty bands.

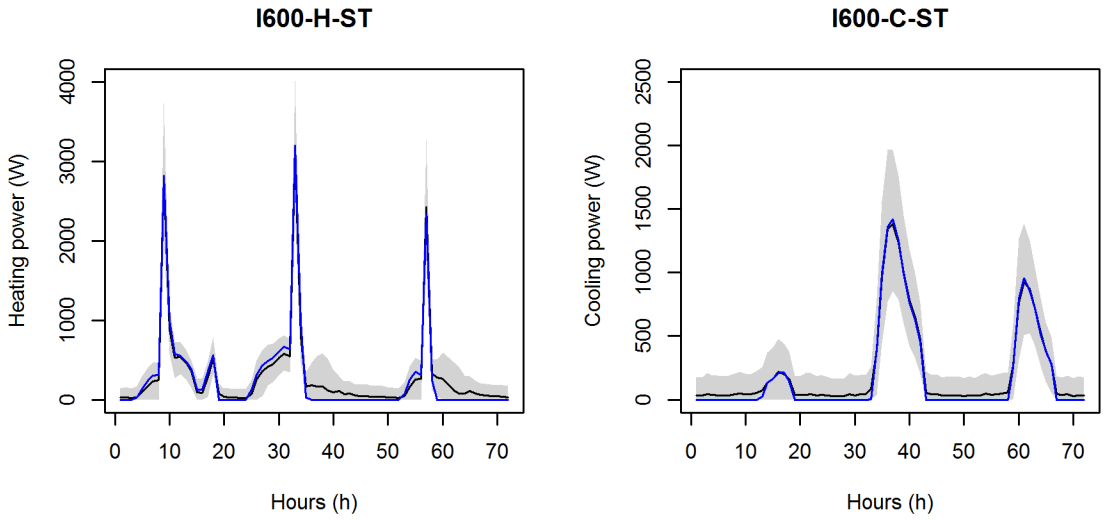


Figure 3.2 Example result of baseline model with 95% confidence interval

The accompanying sensitivity analysis uses the Morris method (Morris, 1991) to identify important parameters responsible for the output variations. Menberg et al. (2016) provided a detailed explanation of the procedure and underlying logic. This case study constructed a 10-level design of 148 parameters with a sample size of 1490. The difference between predictions and observations averaged over the three-day period of each test is used as the response under analysis. The absolute mean μ^*, i.e. the expected change in response due to variation in each parameter, is used to rank parameters. The standard deviation of those changes of an individual parameter σ reflects the magnitude of its non-linear effects and interactions with other parameters. Figure 3.3 plots the $\mu^* - \sigma$ plots of each output, in which parameters appearing on the right have large main effects on the output and those appearing on the top have large non-linear effects and interactions. The

results clearly show that coefficients of stack effect in envelope convection, room air temperature gradient, return air offset, ground reflectance, and glazing solar transmittance contribute the most to output variations. The detailed result in terms of the top 20 parameters as well as the complete figure can be found in Table A.1-Table A.6 and Figure A.2 in APPENDIX A.

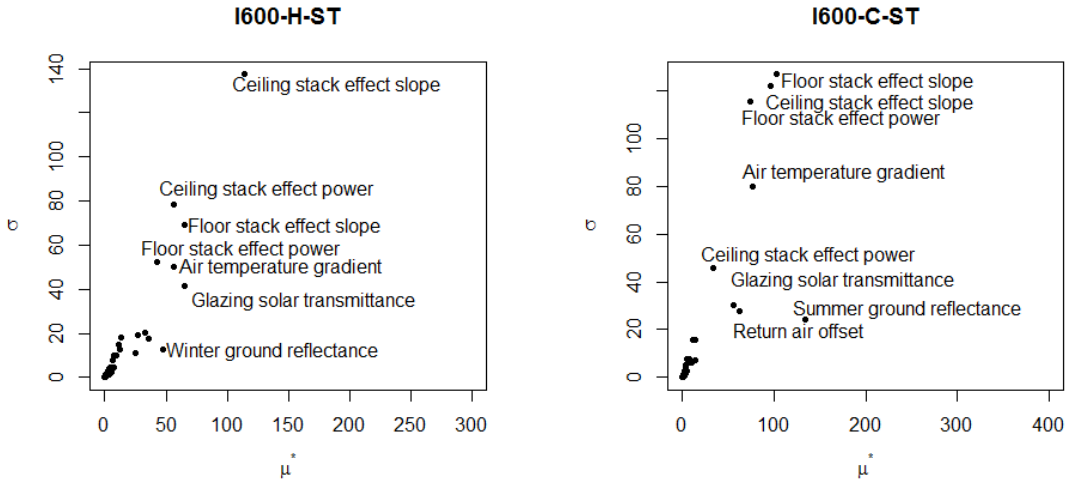


Figure 3.3 Example result of sensitivity analysis of baseline internal model

3.5 Uncertainty propagation using detailed measurements

This section uses as examples interior surface convection, room air temperature gradient, glazing transmittance, and ground reflectance to demonstrate the process of model calibration via uncertainty propagation using detailed measurements. The impact of this process on the refined internal model prediction is evaluated. Return air offset can be estimated in a similar way by measuring the return air temperature and calculating its difference with mean room air temperature at each time step.

3.5.1 Stack effect coefficients for envelope convection

A heat balance equation on an interior envelope surface can be expressed as:

$$R_{ir} + R_{sw} + Q_{cond} + Q_{conv} = 0 \tag{15}$$

where R_{ir} is infrared radiation heat gain, R_{sw} is shortwave (including solar) radiation heat gain, Q_{cond} and Q_{conv} are surface conduction and convection heat gains. For Q_{conv}:

$$Q_{conv} = h_c A\left(T_{air} - T_{surf}\right) \tag{16}$$

$$h_c = \alpha\left|T_{air} - T_{surf}\right|^n, log(h_c) = log\alpha + nlog\left|T_{air} - T_{surf}\right| \tag{17}$$

Therefore, by measuring incident infrared and shortwave radiation, material absorptance, surface conduction heat flux, and surface and adjacent air temperature, the convective heat transfer coefficient at each time step can be calculated, and a simple linear regression to temperature difference can be performed to estimate the two coefficients. Table 3.4 shows the estimates and standard deviation for wall, ceiling, and floor, along with their respective R^2 of the linear regression. In the following simulation, the regression coefficients and residuals will be independently sampled and summed to calculate the instance heat transfer coefficients.

Table 3.4 Coefficient estimates from simple linear regression

Envelop	Coefficient	Estimate	Std. Error	R^2
Wall	$log\alpha$	0.2245	0.0122	0.5043
	n	0.3182	0.0108	
Ceiling	$log\alpha$	-0.1853	0.0111	0.4363
	n	0.3295	0.0127	
Floor	$log\alpha$	0.3532	0.0093	0.3376
	n	0.2556	0.0122	

3.5.2 Room air temperature gradient

Air temperature measurements close to ceiling and floor are used to estimate the presumed constant room air temperature gradient, assuming estimates from measurements on each hour, as shown in Equation 18, are independently and identical sample points from the same underlying distribution plus the aggregated sensor error due to subtraction of two temperature sensor readings:

$$\hat{G}_i = \frac{T_i^{ceiling} - T_i^{floor}}{h} \tag{18}$$

where $\hat{G}_i$ is the estimates at time i, $i = 1,2,\dots t$, $T_i^{ceiling}$ and T_i^{floor} are corresponding temperature readings subject to sensor error, and h is the room height. Result using synthetic measurements is shown in Figure 3.4, indicating the uncertainty is apparently constrained from $\mathcal{U}(0.5,1.875)$ and covering the true value 0.7023. The estimates are used as an empirical sample from which points will be drawn and used in the refined simulations.

3.5.3 Glazing solar transmittance

Similarly, estimates of the single glazing solar transmittance comes from the division of window transmitted and incident solar insolation from pyranometer readings plus sensor errors. However, results in Figure 3.4 show a skewed and indeed biased distribution from the true value 0.7812, probably because of the effects of window frames and dividers. This suggests that further measures are needed to improve the uncertainty quantification method.

3.5.4 Ground reflectance

Ground reflectance in summer and winter are estimated by measuring the ground global horizontal radiation from the on-site weather station, and ground reflected component on the south-facing vertical incident insolation. The estimates pooled from hourly readings as shown also in Figure 3.4 well cover the true value 0.2371 and 0.3370 respectively, and will be fed directly into the refined model.

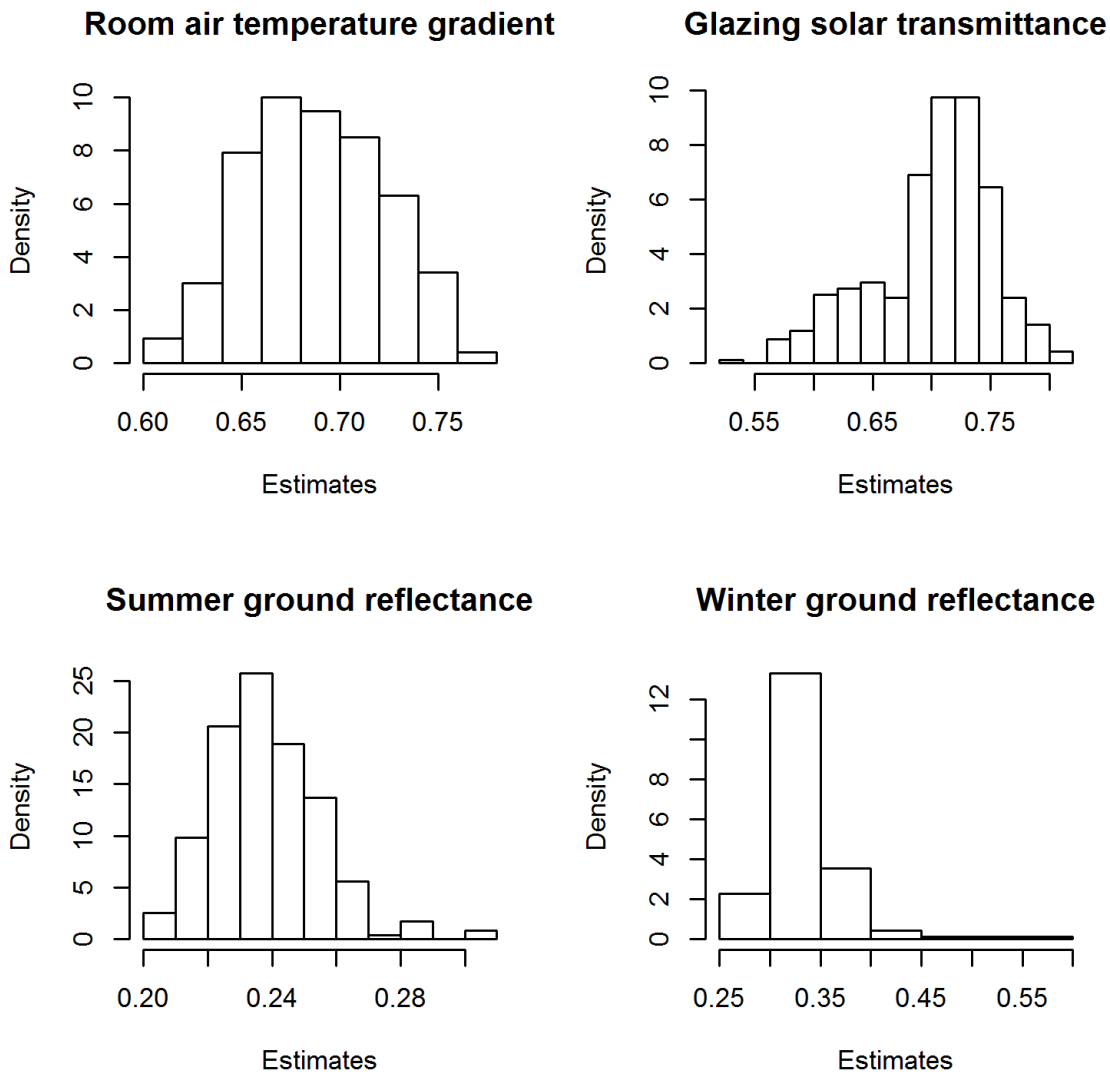


Figure 3.4 Distribution of refined parameter estimates

3.5.5 Refined model analysis result

The same LHD sample is applied to the calibrated internal model with refined quantification of uncertainty in the above parameters except glazing solar transmittance. Example results of the refined probabilistic prediction in Figure 3.5 (pink) show apparent improvement on constraining the output uncertainties, proving the effectiveness of model calibration via uncertainty propagation using detailed measurements. However, apparent

deviations from the measurement persist because of the free-floating phenomenon during AHU heating test and the sensitivity of heating power to setpoint noise, which cannot be eliminated even using the "true" model with the measured cell air temperature as setpoint in a test run. This indicates that a better designed indoor condition might be needed for the experiment to be performed, or the period when air-conditioning is off is excluded in the assessment. The full results are shown in Figure A.3 in APPENDIX A. The analysis result tentatively accepts the first research hypothesis: the proposed calibration method, particularly uncertainty propagation using detailed measurements, makes improved use of information and data in constraining uncertainty and improving prediction. Quantitative interpretation of the result using the CRPS and associated validation criteria will be discussed in the following section.

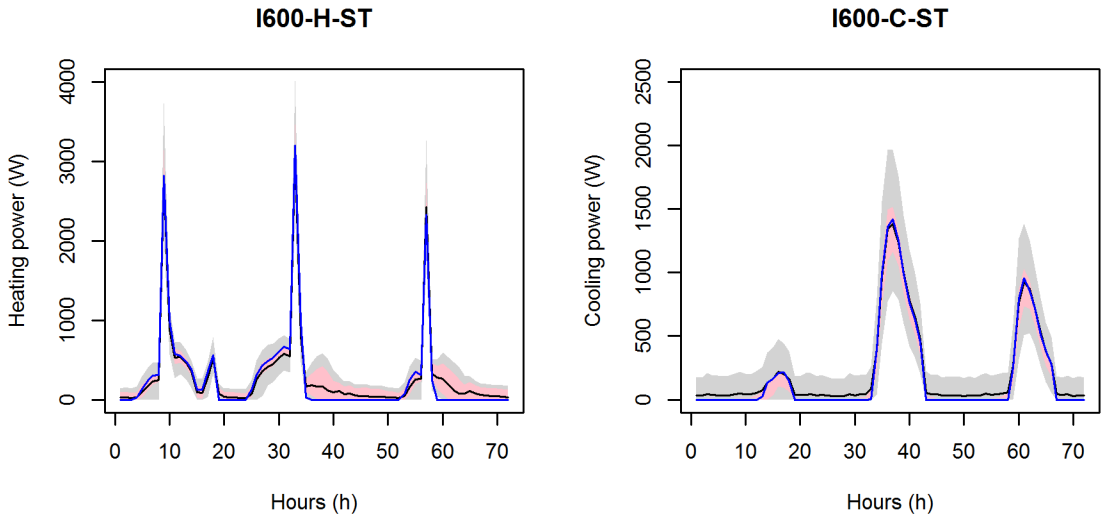


Figure 3.5 Example result of refined model with 95% confidence interval

3.6 Empirical validation criteria

3.6.1 Internal model prediction accuracy under accuracy metrics

The Results of quantitative assessment, using both the CV-RMSE and the nMCRPS on heating/cooling power and their un-normalized counterparts on cell mean air temperature, are shown in Figure 3.6. Since the CV-RMSE only uses the mean prediction, which is barely changed in the refined internal model, this metric does not adequately reflect those refinements. In contrast, results using the nMCRPS show apparent improvement except in two heating tests, where the heating power's sensitivity to setpoint perturbations causes lingering deviations from the actual zero heating power.

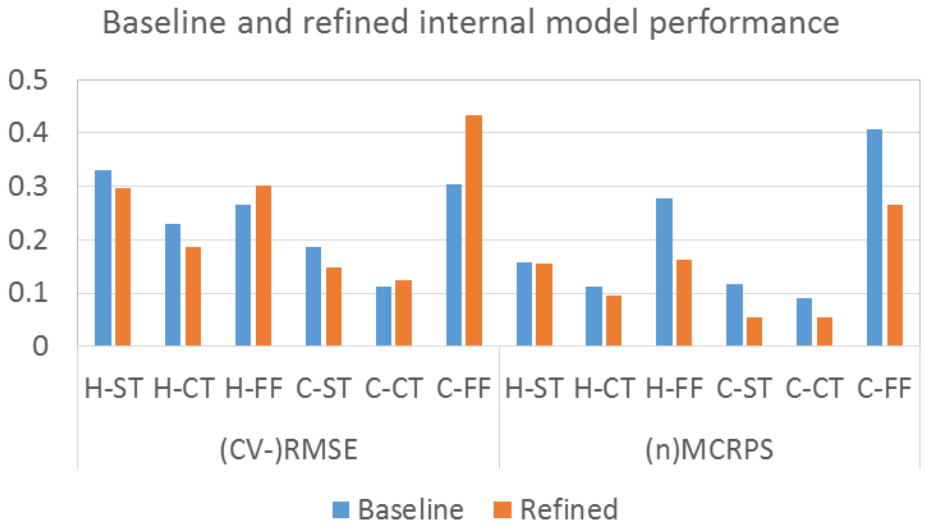


Figure 3.6 Internal model accuracy under different metrics

3.6.2 Validation criteria and validation risk assessment

The probabilistic prediction of the internal model is believed to agree well with observations, after accounting for various parameter and model form errors and propagating associated uncertainties using the available information and data. In this sense, one would expect the discrepancy of an external model prediction to fall within the range of the internal model's if this external model possesses reasonable parameter and model

form errors, and conversely an external prediction with error exceeding this range is probably subject to inappropriate model assumptions, modelling mistakes, computation errors, etc. Therefore, a distribution of metric values regarding each instance as a single prediction is generated, and for example using its 90% quantile as the validation criteria threshold reflects the belief that a truly valid model only has a 10% chance of being rejected. This indicates a 10% risk of the *type I error*, or the *false positive rate* in statistical hypothesis testing.

As the nMCRPS reduces to the normalized mean absolute error (nMAE) for point (deterministic) prediction, the 90% quantile of the nMAE in each test is shown in Table 3.5, which also reflects the improvement due to uncertainty propagation. The baseline and refined distribution of the nMAE of the internal model in all the six tests can be found in Figure A.4 and Figure A.5 in APPENDIX A. While the risk of type I error is directly controlled by the quantile probability, the risk of the *type II error/false negative rate*, i.e. the error of accepting an invalid prediction, depends also on a distribution of errors of invalid predictions. At this stage, if treating the baseline distribution as the set of invalid predictions to be rejected, a hypothetical speculation on the risk of the type II error shows a decrease from 90% to for example 20.85% in I600-C-ST test. Table 3.5 also summarizes the reduced risks in all the six tests. These results also show that the nMCRPS is a more informative accuracy metric than the CV-RMSE for probabilistic predictions in terms of the associated risks. The reduced risks in turn quantifies the informativeness of the detailed measurements regarding the constraining of important parameter uncertainties, and can be used directly in a risk-conscious design of empirical validation experiments for choosing among alternative plans. Therefore, the results of this case study tentatively accepts the

second research hypothesis, proving the framework's effectiveness in representing model validity and data informativeness under uncertainty.

Table 3.5 Baseline and refined validation threshold and reduced risk

Test	Baseline 90% quantile	Refined 90% quantile	Reduced risk
I600-H-ST	0.54	0.33	36.65%
I600-H-CT	0.37	0.19	52.80%
I600-H-FF	2.05	0.58	46.15%
I600-C-ST	0.70	0.18	69.15%
I600-C-CT	0.49	0.15	65.25%
I600-C-FF	2.68	0.91	48.85%

It is worth noting that these type II error risks are calculated based on a hypothetical representation of potential invalid predictions. A more accurate representation would involve extensive investigations on a variety of model form errors, and potentially systematic quantification of modeler's bias via human-subject studies. Nevertheless, these studies are beyond the scope of this dissertation because of the constraints in time and resource, and are expected to be addressed in future work.

CHAPTER 4.

HYDRONIC HEATING SYSTEM INTERVENTION ANALYSIS

In contrast to uncertainty propagation, inverse modelling using Bayesian inference quantifies uncertainty of model inputs via maximizing those inputs' posterior probability that consists of prior knowledge and observation likelihood conditioned on those inputs. This approach makes direct use of often readily available system output observations and explicit formulation of human knowledge, so typically requires much less time and effort than forward uncertainty quantification. Nevertheless, the coverage, representativeness, and quality of observations become vital for a successful and valid calibration of a BPS model especially of a complex system, and this requires extensive studies on a variety of building performance management practices. In the second case study, this dissertation chooses a hypothetical intervention analysis of an existing building with hydronic heating on Cambridge, UK campus as an example of inverse calibration under real-world uncontrolled usage conditions. In addition to the effectiveness of Bayesian inference, this case study will assess the informativeness of monitoring data varying in temporal and categorical scales to test the framework and provide practical insights.

4.1 Building description

The 4-floor case study building is located on University of Cambridge campus in UK. The majority of the building, consisting of cellular and open offices, meeting and seminar rooms, and stairs and corridors, is mainly conditioned by a radiator-based varying-temperature (VT) hydronic heating system. In addition, a constant-temperature (CT)

underfloor heating system is used for the lobby, and two air handling units (AHUs) provide both heating and cooling to address the extra internal load and ventilation demand in a few meeting and seminar rooms, and to satisfy strict thermal condition requirements of the building's paper and film archives. The heating source consists of two cycling gas boilers, and the cooling comes from an air-cooled chiller. A supplementary ground source heat pump (GSHP) serves to pre-heat or -cool the return water.

This case study focuses on modelling and calibrating the south part of the first floor of the building for demonstration purpose, which has its own lighting and plug load consumption metered at 15-min intervals. This part of the building consists of 22 rooms, mostly cellular offices, that represents the typical composition of the building, and is conditioned by two local heating loops that have their total consumption metered incrementally also at 15-min intervals (Figure A.6 in APPENDIX A). The meeting room has a sensor monitoring room air temperature at the same time resolution.

Figure 4.1 Case study building overview

4.2 Model development

4.2.1 Background

The conventional way to model energy supply systems uses an aggregated and constant system efficiency value for the entire system. This approach ignores individual system's dynamic interactions with the building fabric under changes in weather conditions and occupant behavior, which prevents identifying possible causes of system underperformance, and impairs prediction of future interventions with confidence.

Unlike air systems, hydronic heating using radiators as heat emitters has a set of distinctive characteristics. First, each individual radiator is controlled by a thermostatic radiator valve (TRV), which relates hot water flow rate to room temperature and user setpoint via a preset characteristic curve. Therefore, different from an explicit temperature maintainer in air systems, the actual balance temperature of a radiator-heated room depends not only on user settings and system capability, but also on other room conditions and thermal processes. Hence, a high-fidelity model should fully couple room and radiator in modelling the related heat transfer phenomena. Second, since each radiator can be individually controlled, the common way to model buildings with central air systems would be inappropriate, as aggregating similar individual rooms into large thermal zones may lead to errors when a large usage variability is present. Therefore, instead of using existing models in current simulation tools, this case study develops a novel numerical model based on fundamental principles of heat transfer and building physics, and explores the effective model form in supporting the hypothetical intervention analysis.

4.2.2 Modelling methodology

This dissertation proposes a high fidelity three-level modelling method that enables modelling all the thermal processes occurring in individual rooms while efficiently modelling the whole building to estimate heating system performance. Figure 4.2 shows a schematic overview of this method.

First, in each individual room, the room and the radiator are coupled in a full state-space model that captures all the heat transfer phenomena in the room, including convective and radiant heat transfer with the radiator. The radiator, along with the hot water flowing through it, is discretized into a set of elements to represent the temperature decrease along the flow. Following the same approach to representing the TRV characteristic curve as from Xu et al. (2008), the hot water flow rate is modelled as a bounded second-order polynomial function of the difference between room air temperature and the setpoint.

Then, in scaling up to the whole building, while a model with highest fidelity would model each individual room explicitly, this method proposes a simplified prototype-room method that assembles the prototype-building approach in building stock modelling (Deru et al., 2011). For a group of rooms sharing similar typological features and usage patterns, instead of aggregating them into a single large thermal zone in common practices, this method only models a single room explicitly to represent the whole group. A symmetric heat transfer transformation is applied to the full state-space model regarding inter-room heat transfer through partition walls. This transformation assumes the modelled room is located in an infinite array of exactly the same room adjacent to each other horizontally and/or vertically ("Whole building" in Figure 4.2), so for example the floor surface of the modelled room always possesses the same status as the floor surface of the room above,

and equivalently becomes the boundary condition of the ceiling of the modelled room. Therefore, by connecting the floor and ceiling nodes of this single room in the state-space model, the correct heat balance of the room is maintained without modelling the other rooms, and multipliers can be used to obtain the total system output of the whole array. This transformation can be easily combined with explicit modelling of inter-room heat transfer through partition walls in modelling the entire building with a few partitions being assumed adiabatic, so the entire building becomes a combination of all these groups of rooms, each represented by a single prototype room. This method maintains not only the granularity in depicting heat transfer and heat balance in individual rooms, but also a correct representation of the whole building's thermal behavior in interacting with the heating supply system.

Finally, because of the relatively small time constant of the supply system, this approach specifies the boundary conditions of the state-space model based on actual operation schedules and weather and usage conditions, and adopts steady-state equations from BS EN 15316-2-3:2007 standard to characterize thermal loss and auxiliary energy. The inlet water temperature and flow rate of each individual radiator are determined by the central supply water control logic and the thermal loss along the pipes, whereas the outlet water temperature and flow rate are determined by room heat transfer process, solved by forward Euler method of finite difference discretization, and used as the inputs of the supply system model.

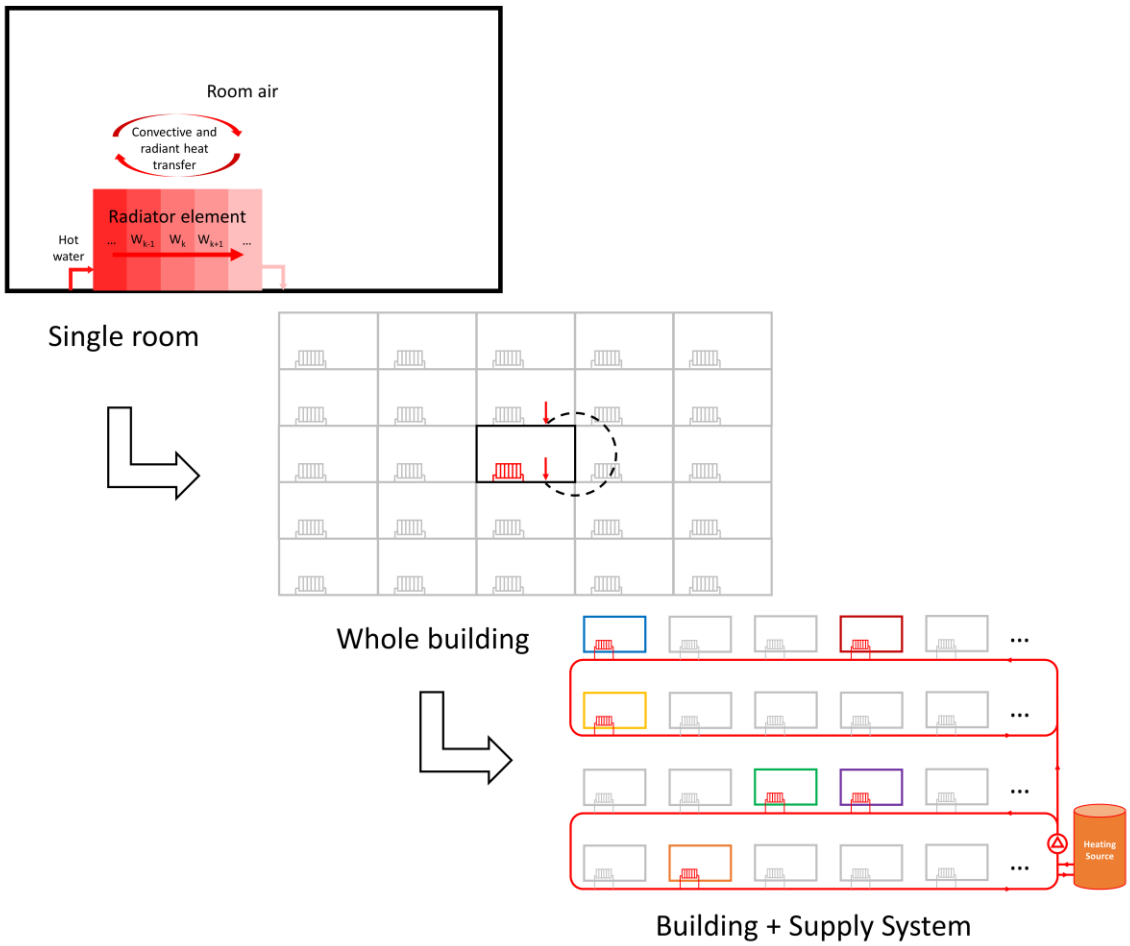


Figure 4.2 Schematic overview of the three-level modelling method

4.2.3 Model effectiveness

To determine if the proposed modelling method, especially regarding the prototype-room approach, satisfies the need of the hypothetical intervention analysis, an assessment on the impact of the set of modelling simplifications is performed under a variety of building, weather, and usage uncertainties. The models under assessment are as follows:

1) Baseline: a highest-fidelity full state-space model that models every room of the building individually, and considers usage variabilities, including

occupancy, lighting and plug load use, window opening, thermostat setting, of individual rooms of the same room type, e.g. cellular offices, meeting rooms, etc.

2) Model 0 (M0): a full state-space model that models every room of the building individually, and considers usage variabilities only between rooms of different room types.
3) Model 1 (M1): a full state-space model that models every room of the building individually albeit using prototype-room approach, i.e. rooms of the same prototype are modelled the same but simulated separately plus appropriate heat transfer transformations.
4) Model 2 (M2): a full state-space model that applies the complete prototype-room approach, i.e. only one single room is modelled and simulated to represent each group.

A variety of uncertainties associated with the modelling and simulation becomes external uncertainties, equivalently scenario uncertainties when different model forms are compared. The basis of parameter uncertainty associated with the building fabric comes primarily from the literature (Macdonald, 2002; Sun, 2014; Heo et al., 2015; Wang, 2016). Parameters including peak load density, base load density, and length of peak load period are used to parameterize occupancy profiles, and the uncertainties of these parameters are quantified based on standards and experiences. Uncertainties in lighting and plug load adopts the modelling approach proposed by Ward et al. (2017), which builds upon functional principal component analysis (FPCA) of internal load usage in another campus building with similar functions. The usage differs by room functions and

weekday/weekend. Modelling of window opening behavior adopts the model used by Rijal et al. (2007), which regresses the probability of window opening to indoor and outdoor conditions. Nevertheless, it is assumed that the window will only be opened if this probability exceeds 50% to ensure consistency in comparing the different models. The shading screens are assumed to be closed once the incident solar radiation exceeds an unknown threshold that is uniform across all the windows. Because of lack of literature, uncertainty of the hydronic heating system is quantified based primarily on building specifications and experience judgement. Uncertainties in weather conditions are represented by variabilities in individual days/hours. Uncertainty in effective leakage area per exterior envelope area adopts the $lognormal(1.28, 0.88^2)$ distribution from the work of Wang (2016), whereas the rest uses beta distributions instead of the more commonly used triangle distributions because of its smooth and continuous probability density functions. The beta distribution is also preferred than a truncated normal distribution because of the former's flatter and skewed probability density function. A complete summary of parameter uncertainties is shown in Table A.7 in APPENDIX A. In addition, a perturbation following a normal distribution with a standard deviation of 5% is added to occupancy, internal load, and window opening area of each room to represent individual variabilities in the baseline model. A normal distribution with a standard deviation of 0.5°C is applied to their TRV settings for the same purpose.

A LHD sample of 65 parameters with a sample size of 2520 is constructed, each assigned a single day of the actual weather conditions, weekday/weekend status, heating on/off schedules, and VT loop hot water supply temperature from a 5-week period from 02/20/2017-03/26/2017. Each simulation also includes the previous day as warm-up to

reduce the impact of initialization. The system outputs of interest includes daily/hourly total heating energy consumption in kWh, comprised of both room heating consumptions and the local loop thermal loss, and meeting room daily/hourly mean air temperature. Figure 4.3 shows the deviation from each model to the highest-fidelity baseline model, where heating consumptions are also normalized by the respective baseline mean consumptions. The results suggest that the effect of prototype-room modelling approach on total heating consumption is negligible compared to the ignorance of individual room variabilities, and the error magnitude of the latter is in general less than 5% but could be as much as 30% in daily output and 100% in hourly output. In contrast, the prototype-room approach tends to underestimate both daily and hourly mean room air temperature, although the error magnitude remains relatively small. Ignorance of individual room variabilities still accounts for most of the errors, with overall 0.1°C off the baseline but could be as much as 0.3°C in daily mean room air temperature and 0.8°C in hourly output.

An explorative multiple linear regression between the M2 daily error and the uncertain parameters reveals that parameters including radiator characteristics, glazing equivalent U-value, effective leakage area, shading control threshold, and cellular room thermal mass and setpoint appear to be significant at 5% level, indicating their strong linear correlation with the errors. However, they only explain 4.42% of the variations of total heating consumption errors and 19.63% of variations of average temperature errors, suggesting a negligible effect. Therefore, it is concluded that the proposed prototype-room modelling approach is effective in terms of representing the thermal process in the building under study. The M2 model will be used in the following calibration process as the physical model, and it is expected that the model form error due to the prototype-room approach,

mostly uncorrelated with uncertain model parameters, will be accounted by the statistical model bias and random error terms in the Bayesian calibration framework.

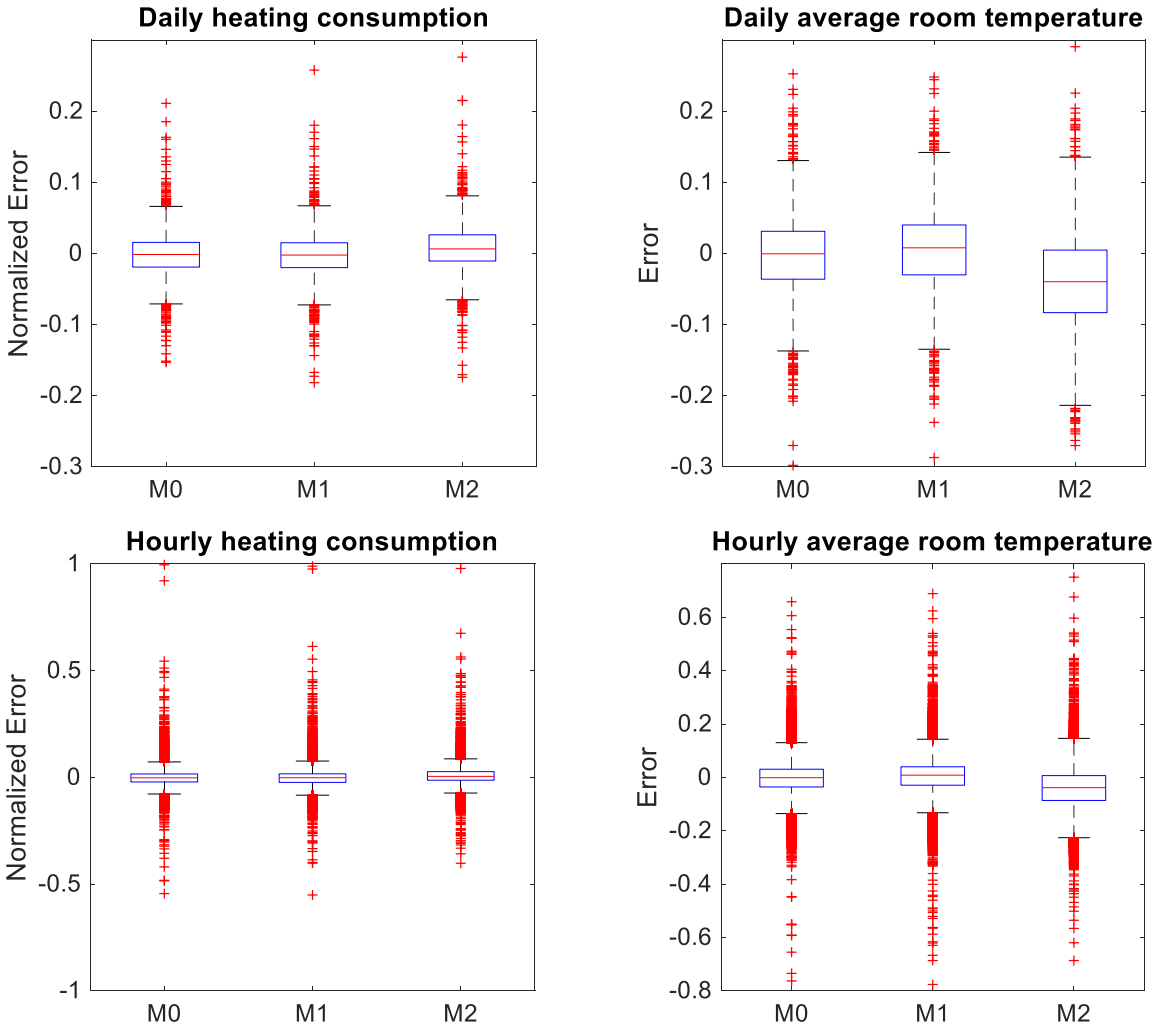


Figure 4.3 Error of system outputs of three simplified models

4.3 Model calibration

4.3.1 The Bayesian calibration framework

This dissertation adopts the classic Bayesian calibration framework of computer models from Kennedy and O'Hagan (2001) as shown in Equation 19:

$$y = \eta(x, t) + \delta(x) + \epsilon \tag{19}$$

where y is the field observations, usually standardized with zero mean and unit standard deviation to ensure that all the types of observations are of the same magnitude and considered equally important. $\eta(x, t)$ is the outputs of the physical model, represented as a function of variable inputs x, usually known and varying during the observation of system outputs, and calibration parameters t, unknown but fixed building features. This framework also considers model form error by including the term $\delta(x)$, assumed to only depend on variable inputs, and random observation error ϵ, usually assumed to follow a Gaussian distribution with an unknown variance, i.e. $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

This classic framework formulates $\eta(x, t)$ and $\delta(x)$ as two kriging models with Gaussian kernel, equivalently two Gaussian process models. This case study retains these formulations and takes the modular-maximum likelihood estimate (modular-MLE) approach proposed by Bayarri et al. (2007) as a more efficient alternative to a full Bayesian analysis where possible. Regarding $\eta(x, t)$, while Li et al. (2015) proposed to use the exact physical model or a polynomial regression emulator, in this case study the former is prohibitively expensive in computation and the latter did not render a good approximation of the complex physical model in early explorations. As for $\delta(x)$, Tuo and Wu (2016) pointed out that this formulation is inconsistent in estimating calibration parameters because of its dependence on specific covariance functions, and therefore would impair the calibration credibility. The authors instead proposed the definition of least L_2 distance calibration, which defines the true value of calibration parameters to be those minimizing the L_2 norm of the observed discrepancy. This definition provides a well-defined

calibration problem on a theoretical basis, resolves the identifiability of calibration parameters and model form error, and increases the estimation efficiency. A Bayesian version of this framework can be found in Plumlee (2016). However, its implementation involves evaluating the gradient of the log-likelihood function to the calibration parameters, which adds to the already expensive HMC computation. In addition, the calibration problem in BPS applications is arguably well defined by the underlying physical phenomena. As model form error is not necessarily unbiased, the true model parameter values, which correspond to certain physical properties, may not coincide with those that minimize the L_2 norm of the discrepancy. More importantly, a pragmatic view of statistical model calibration from the perspective of engineering practices regards the calibrated physical model as the intended deliverable, and treats the explicit formulation of model bias in the meta-model as more of a way to decompose sources of error and prevent over-fitting. Without precisely knowing the form and magnitude of model form error a priori, the classic Bayesian calibration framework appears to be more favorable because of its model agnostic assumption and special flexibility in capturing non-linear patterns. Nevertheless, as Higdon et al. (2004) suggests, a modeler should be cautious about the calibration result, and revisit the physical model if a large model form error is present. Only this can ensure that the physical model is reliable for later applications where considerable extrapolations may occur. The framework of Tuo and Wu (2016) is suggested for future work for its more elegant definition, better estimation performance, and potentially more effective calibration of physical models.

4.3.2 Calibration scenarios

To assess the informativeness of different monitoring data and in turn test the proposed framework, this case study performs Bayesian calibration of the physical in six scenarios where the available observations vary in categorical and temporal scales. A summary of the six scenarios is shown in Table 4.1.

Table 4.1 Summary of calibration sceanrios

Scenario	Available data categories	Resolution
D-BI	Heating, electricity, meeting room air temperature	Daily
D-UI	Heating and electricity	Daily
D-EI	Heating only	Daily
H-BI	Heating, electricity, meeting room air temperature	Hourly
H-UI	Heating and electricity	Hourly
H-EI	Heating only	Hourly

4.3.3 Parameter screening

To maintain the inverse modelling problem tractable, only a few important model parameters are typically calibrated. Parameter screening using the Morris method is therefore performed to select those parameters to be included in the Gaussian process emulator and calibrated using Bayesian inference. A 30-level design of 65 parameters with a sample size of 1980 is applied to the physical model, with the absolute errors toward field observations for the daily and hourly heating consumption and meeting room mean air temperature as the system outputs under analysis. The 65 parameters include a day sampler linking to actual weather conditions and heating system operations to be used for the single day simulation, and an hour sampler specifying the output of which hour of that day is used as the output of analysis. Figure 4.4 shows that overall the daily and hourly outputs share similar important parameters, including radiator characteristics, infiltration, glazing solar

heat gain, room occupancy and usage, and day and hour samplers. The detailed result in terms of the top 20 parameters as well as a larger figure can be found in Table A.8-Table A.11 and Figure A.7 in APPENDIX A.

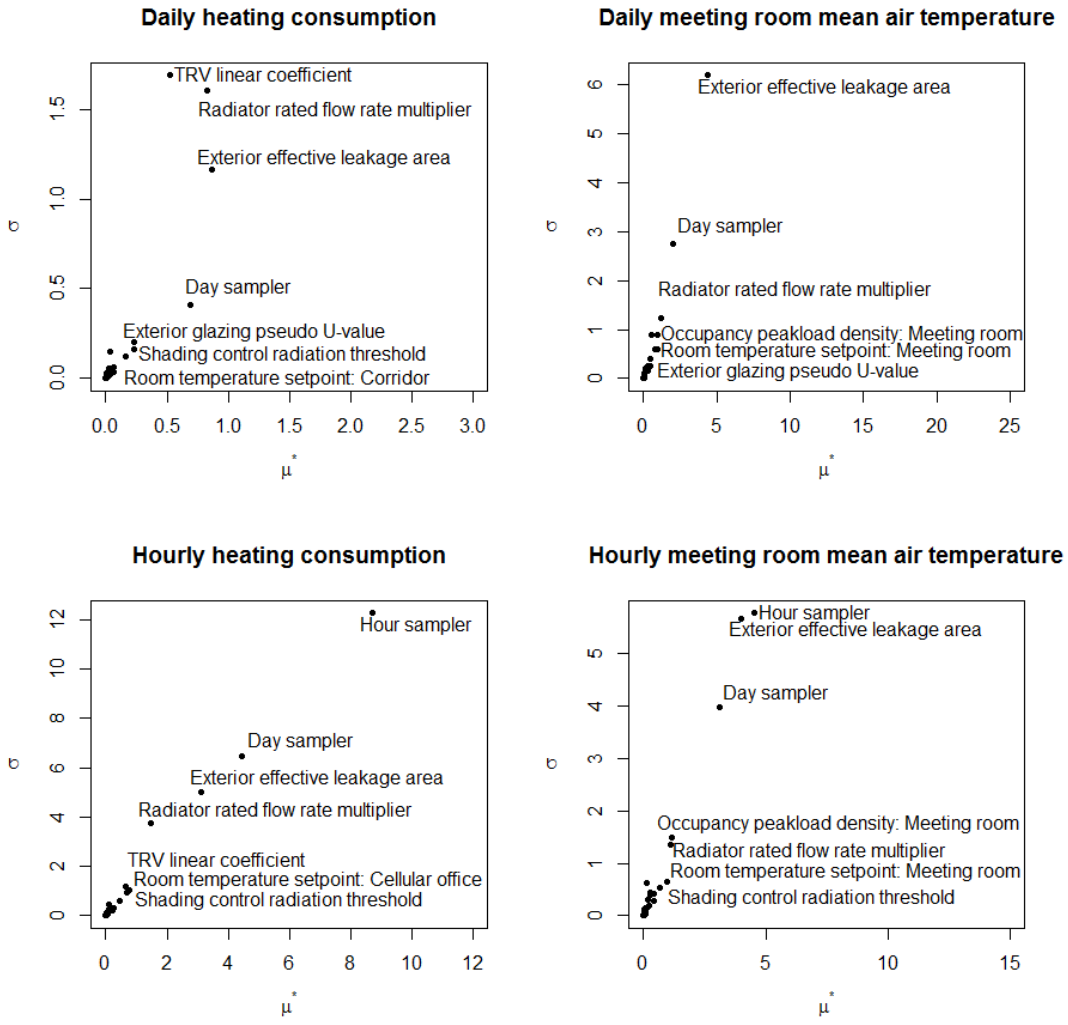


Figure 4.4 Result of parameter screening

Table 4.2 summarizes the model inputs to be included in the emulator in each calibration scenario. Calibration parameters are chosen based on the above parameter screening results. Variable inputs, including weather and usage conditions and heating system operations, are confounded by the real conditions used in the simulation. Therefore,

these inputs cannot be analyzed separately in the parameter screening, and their selection are solely based on previous experience. Since the modelling of internal load uses samples of random sequences from Ward (2017) instead of a parametric model, the daily/hourly total electricity consumption is used as a feature in the analysis, and will serve as either a variable input or a calibration parameter in the following Bayesian calibration depending on the availability of electricity monitoring. In the meantime, the output type indicator is used as a dummy variable such that standardized heating consumption and room temperature outputs can be pooled together to train a single model. This avoids the difficulty in combining the parameter estimation results of two separate calibrations. A similar approach has been used in a polynomial regression model in the work of Li et al. (2016), which also includes a more detailed explanation about this technique.

Figure 4.5 provides an overview of all the three types of filed observations. It is noticed that the peak consumption appears one or two hours after the heating is turned on, so the heating on/off status on the hour before previous hour is chosen as an indicator of this peak load for hourly outputs. Figure 4.5 also shows that the quality of hourly heating consumption data appears to be very poor, probably because of irregular heating operations. At the same time, since heating consumption is accumulated per 15-min and restored to 0 every midnight, and unfortunately the heating meter reading only increases by a fixed interval rather than continuously, the hourly heating consumption, calculated from subtracting the previous reading from the current and then taking the sum of four 15-min consumptions, may not reflect the actual heating consumption even at hourly resolution. Daily consumption is less susceptible to this reading issue as it is the summation

of 72 15-min consumptions. This reflects a potential monitoring robustness issue in real-world building performance management practices.

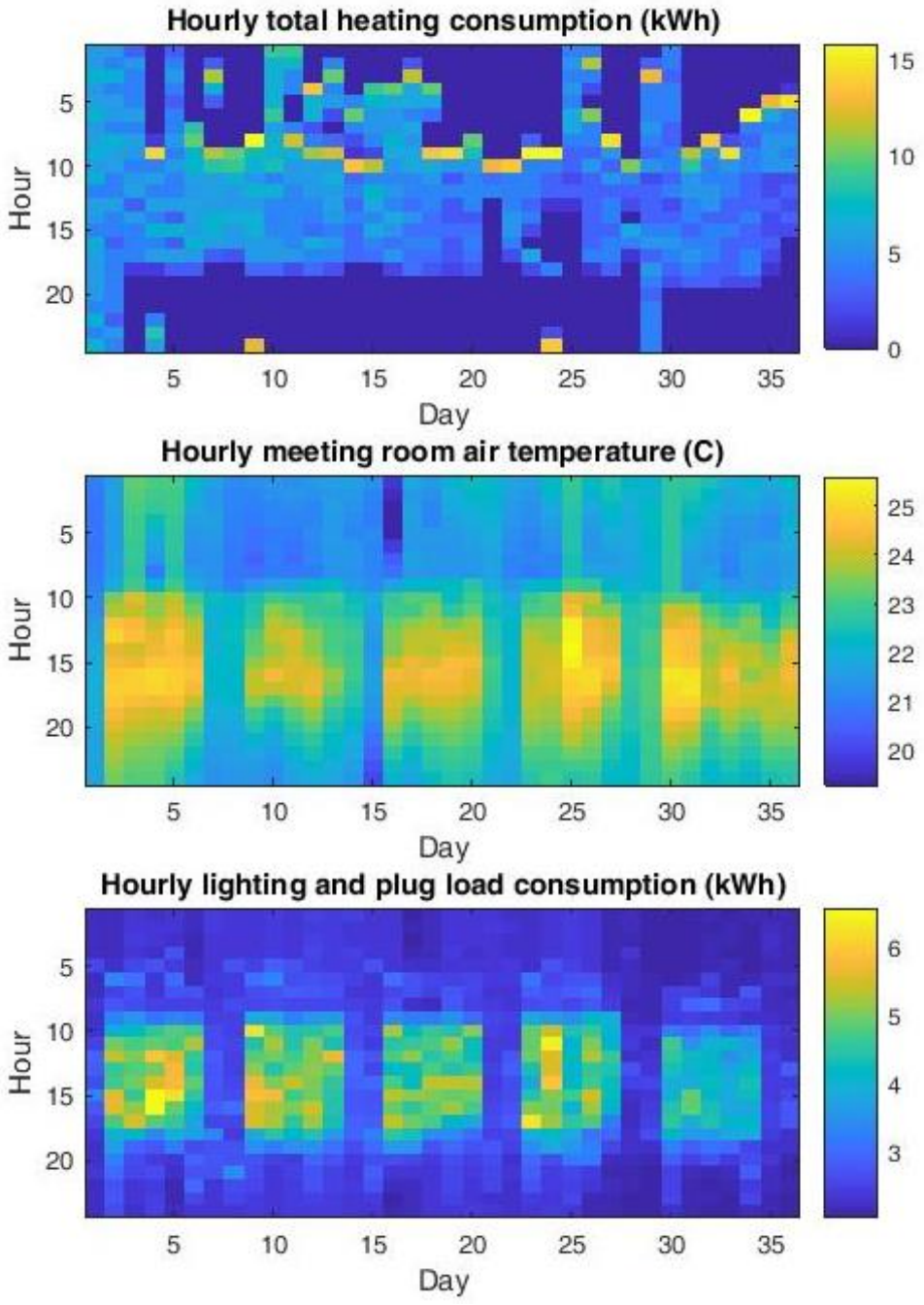


Figure 4.5 Visualization of field observations

Table 4.2 Model inputs in the emualtor in each calibration scenarios

Scenario	Calibration parameters	Variable inputs
D-BI	1,2,3,4,5,6,7,8,9,10,11	1,2,3,4,5,7,9,10
D-UI	1,2,3,4,5,6,7	1,2,3,4,5,7,9
D-EI	1,2,3,4,5,6,7,12	1,2,3,4,5,7
H-BI	1,2,3,4,5,6,7,8,9,10,11	1,2,3,4,6,7,8,9,10
H-UI	1,2,3,4,5,6,7	1,2,3,4,6,7,8,9
H-EI	1,2,3,4,5,6,7,12	1,2,3,4,6,7,8

Index	Input name
Calibration parameter	
1	Radiator area multiplier
2	TRV linear coefficient
3	Radiator rated flow rate multiplier
4	Exterior glazing equivalent U-value
5	Effective leakage area
6	Shading control threshold
7	TRV setpoint: Cellular office
8	Occupancy peak load density: Meeting room
9	Occupancy base load density: Meeting room
10	Occupancy peak load hours: Meeting room
11	TRV setpoint: Meeting room
12	Daily/hourly electricity consumption
Variable inputs	
1	Daily/hourly average dry bulb temperature
2	Daily/hourly average global horizontal radiation
3	Daily/hourly average wind speed
4	Daily/hourly average hot water supply temperature
5	Daily total heating-on hours
6	Heating on/off in the hour before previous hour
7	Weekday/weekend
8	Hour of the day
9	Daily/hourly electricity consumption
10	Output type indicator (heating/temperature)

4.3.4 Physical model emulation

The classic Bayesian calibration framework regards the observed physical model outputs y_c as the true model outputs $\eta(x,t)$ plus computation error ϵ_{en}, i.e. $y_c = \eta(x,t) + \epsilon_{en}$. A Gaussian process emulator of $\eta(x,t) = \eta(d) = GP\left(\mu_\eta, C(\cdot,\cdot)\right)$, where $d = \{x,t\}$, assumes that the true model outputs form a joint multivariate Gaussian distribution determined by mean and covariance matrix in the form of functions of model inputs. Typically, the mean μ_η is assumed to be zero, and the element c_{ij} of the covariance matrix $\Sigma_\eta = C(\cdot,\cdot)$, which represent the covariance between $\eta(d_i)$ and $\eta(d_j)$, takes the squared exponential kernel function:

$$c_{ij} = exp\left(-\frac{1}{\lambda_\eta}\sum_s \beta_\eta^s \left(d_i^s - d_j^s\right)^2\right) \tag{20}$$

where d_i^s is the sth dimension of model inputs, β_η^s is the corresponding weight factor, and λ_η is the overall variance precision. $\phi_\eta = \{\beta_\eta, \lambda_\eta, \epsilon_{en}\}$ is often denoted as hyper-parameters to be distinguished from the physical model inputs d, and is estimated in the emulation. The above formulation serves as a measure of dissimilarity, so outputs are highly correlated when inputs are close enough under this measure. By specifying the structure of the covariance function, a Gaussian process model can flexibly represent the model behavior and obtain an exact fit on the given sample points.

A smaller sample of 840 points is used in the simulation to reduce computation cost in emulation and calibration. Only one hourly output per sample point is chosen by the hour sampler in the hourly emulation for the same reason. Both model inputs and outputs

are standardized by their respective mean and standard deviation. Hourly heating outputs when heating is off are excluded, but the concurrent room temperature outputs remain. 90% of the simulation outputs are used as the training dataset and the rest are held for testing. R package *rstan* is used for the emulation. Maximum likelihood estimation (MLE) is used to estimate hyper-parameters ϕ_η. However, when this approach is not feasible because of the complex shape of the solution space especially in hourly emulation, a full Bayesian analysis using HMC is used instead albeit its considerable computation cost. Testing results using the mean prediction of the Gaussian process emulator with fixed hyper-parameters are shown in Figure 4.6, indicating an overall good daily emulation but relatively poor hourly emulation.

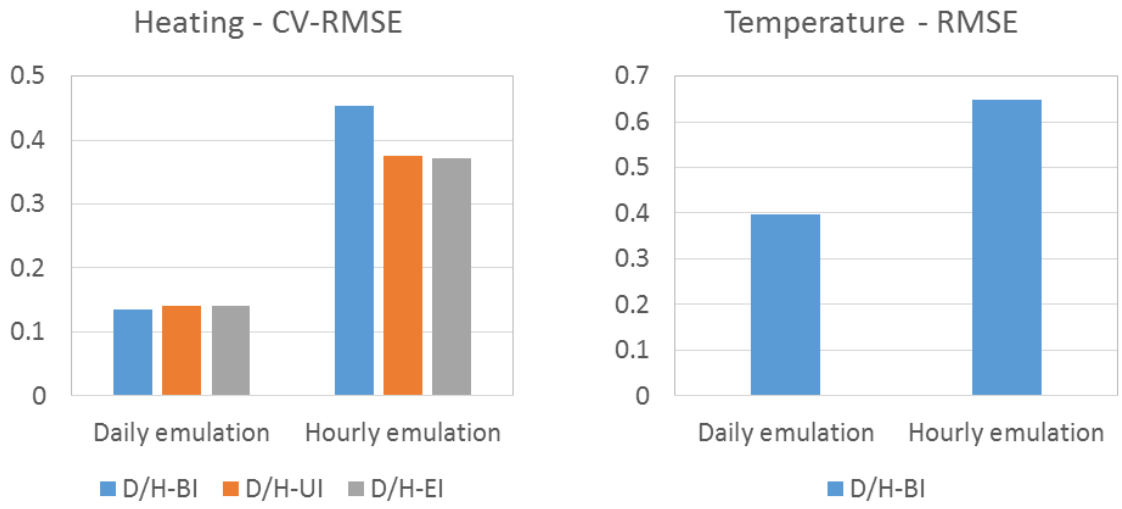


Figure 4.6 Result of emulator testing

4.3.5 Calibration

By formulating the model bias $\delta(x)$ to be another Gaussian process, and denoting by $t = \theta$ the true value of calibration parameters under which field response y is observed,

the calibration problem becomes estimating θ and hyper-parameters $\phi_\delta = \{\beta_\delta, \lambda_\delta, \epsilon_e\}$ of $\delta(x)$. Following Bayes' rule:

$$p(\theta, \phi_\delta|y) \propto p(y|\theta, \phi_\delta)p(\theta)p(\phi_\delta) \tag{21}$$

so the posterior distribution $p(\theta, \phi_\delta|y)$ can be obtained from their prior distributions $p(\theta)$, $p(\phi_\delta)$, and the likelihood of observations:

$$p(y|\theta, \phi_\delta) = |\Sigma|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}[(y-\mu)^T\Sigma^{-1}(y-\mu)]\right\} \tag{22}$$

where Σ is the whole covariance matrix that combines Σ_η and its counterpart in model bias term, Σ_δ, as well as the observation error ϵ_e. μ is the mean vector that is conditioned on the physical model outputs y_c through the covariance determined by the squared exponential function of model inputs. Prior distributions of calibration parameters come from the associated uncertainties, which is the same as being used previously except that the effective leakage area uses $Beta(2.658,10)$ projected to a range of $(0,10)$ instead as this beta distribution keeps the same mode and is bounded to avoid a long tail that complicates the HMC process. Priors of ϕ_δ are assigned based on Guillas et al. (2009) to assume a prior belief of an average 20% model bias and 5% observation error. Observations from the first four weeks, 02/20/17-03/19/17, are used in the calibration, and those from the last week is held as the test dataset.

Assessment using the test dataset includes the predictions not only from the fitted Gaussian process model, but also from the actual physical model by feeding a 100-point subset of the posterior estimates of calibration parameters into the simulation. Since

Bayesian calibration only matches the electricity consumption at an aggregated level, and the related model inputs are indeed sample sequences of lighting and plug load usage, this simulation chooses a simulation instance from the original LHD sample whose electricity consumption matches each of the actual hourly observations, and uses the sample sequences of this instance with further adjustment as the "true" model inputs. This is uniformly applied to all the calibration scenarios, so assessment of the physical model in this sense becomes testing the estimates of calibration parameters.

To assess the effectiveness of Bayesian calibration as compared to deterministic parameter estimation techniques, a calibration using Matlab non-linear optimization function *fmincon* is also performed for each calibration scenario. Only the lower and upper bounds of each calibration parameter are used to define the parameter space, and no model bias is considered. In contrast to Bayesian calibration, electricity consumption is used as model outputs to reflect the common practice. The daily mean profiles of electricity consumption is used as a basis, and adjustment factors are used as calibration parameters to shift these basis profiles to match actual consumptions. The objective function is a weighted sum of squared values of the NMBE and the CV-RMSE as being proposed by Reddy et al. (2007a). This objective function is minimized in the optimization process where *fmincon* searches the parameter space for the optimal values.

Results in Figure 4.7 show that in general Bayesian calibration outperforms its deterministic counterpart in both the statistical and physical models, with an average value of the CV-RMSE of 20% in daily total heating consumption and the RMSE of 0.6°C in meeting room daily mean air temperature, and 50% and 0.8-1.2°C in hourly outputs. This proves its effectiveness in incorporating modeller's experience and observations in the

parameter estimation. The results also suggest that calibration to hourly data is not satisfactory, probably because of relatively poor emulation of the physical model and noisy hourly observations as suggested in Figure 4.5. In summary, the superior performance of Bayesian calibration tentatively accepts the first research hypothesis: the proposed calibration method, particularly Bayesian calibration, makes improved use of information and data in constraining uncertainty and improving prediction.

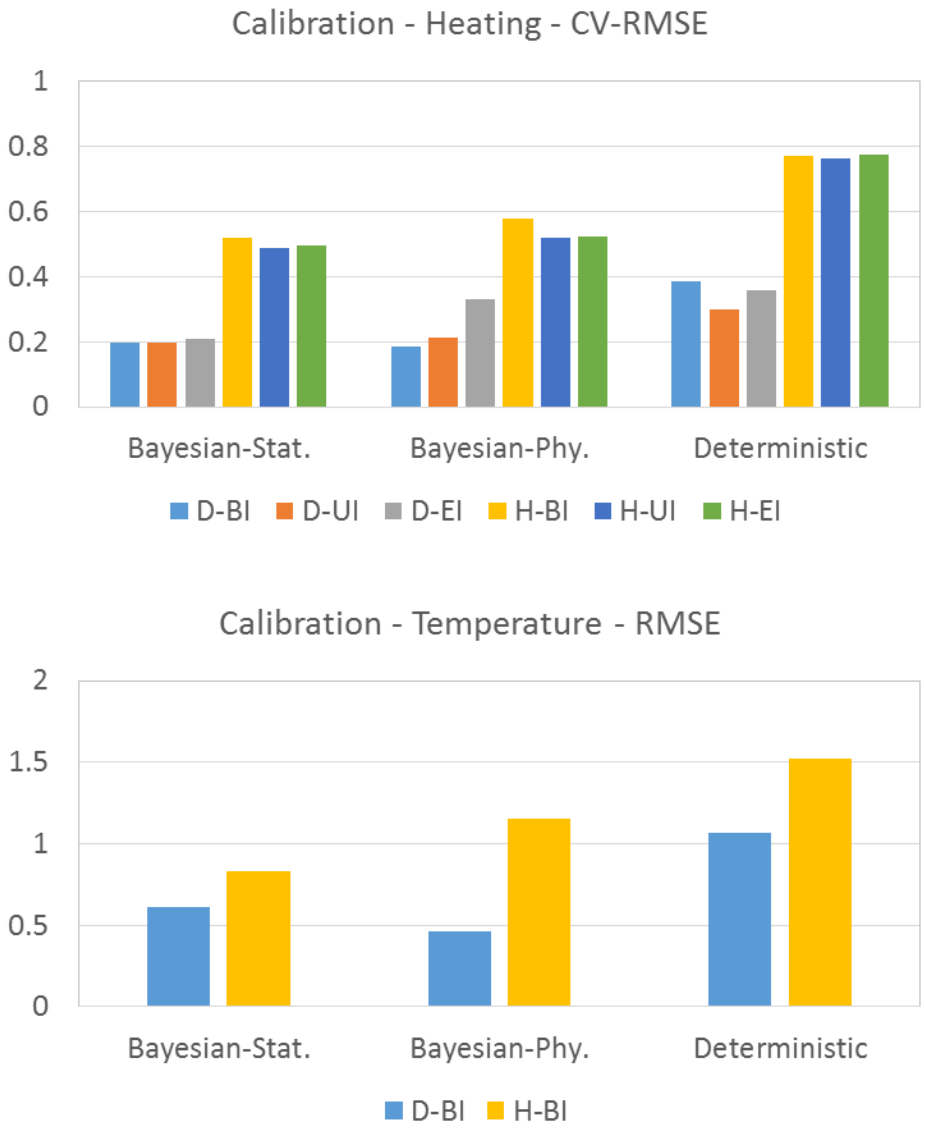


Figure 4.7 Calibration result using Bayesian and deterministic methods

4.4 Risk-conscious building intervention analysis

4.4.1 Physical model accuracy under accuracy metrics

A further assessment of the physical model before and after calibration is performed in this section for the six calibration scenarios. Regarding the calibration parameters, Figure 4.8 plots their prior and posterior estimates in all the six scenarios, with each

calibration parameter identified by the same index as shown in Table 4.2. The detailed comparison of calibration parameter posteriors are shown in APPENDIX A.

The results show an overall consistent estimation of calibration parameters among scenarios with the same types of observations in spite of the difference in the time resolutions. However, outcomes are more disparate for the set of parameters calibrated to different types of observations, suggesting these observations' different informativeness. In addition, only a few of calibration parameters have uncertainty constrained by the observations in each calibration scenario. This suggests the complexity of calibrating a dynamic simulation model to real-world observations, as well as the potential inadequacy of these monitored observations for a definitive understanding of actual building performance. While reducing the number of calibration parameters may help reduce the estimate uncertainty give the same amount of data, it impairs the performance of the Gaussian process emulator and thus the quality of the calibrated physical model. Further improvement on model calibration method is therefore suggested in similar practices.

The results of the models' prediction accuracy on daily and hourly heating consumption and meeting room mean air temperature are shown in Figure 4.9. In general, models are well calibrated to daily observations, but have considerable errors in hourly predictions, exceeding the commonly used 30% threshold for hourly heating consumption and having a minimum 1°C of the RMSE for hourly temperature predictions. This suggests that the information and data used in all the six calibration scenarios are potentially insufficient to support accurate hourly predictions, which coincides with the observations from calibration parameter posteriors.

A closer look at the accuracy of calibrated models among different scenarios shows that, the physical model before calibration has overall good agreement with heating consumption observations, but considerably biased temperature predictions. The physical model after calibration, i.e. refined quantification of uncertainties associated with important parameters and model bias, has the best overall accuracy when all the types of observations are used regardless their time resolutions. In contrast, calibration without electricity consumptions tends to cause erroneous heating and especially temperature predictions; the benefit of temperature monitoring is not significant, probably because its relative small day-to-day variations over the entire period do not contribute too much information in addition to the heating consumption.

In the meantime, comparison between two accuracy metrics shows that the CV-RMSE and the nMCRPS tend to have an overall consistent indication of model accuracy albeit differ in the error magnitudes. This observation suggests that the agreement of the mean prediction (as in the CV-RMSE) alone may serve as a robust model accuracy indicator in inverse modelling applications, where the prediction error mostly comes from estimation bias rather than variances.

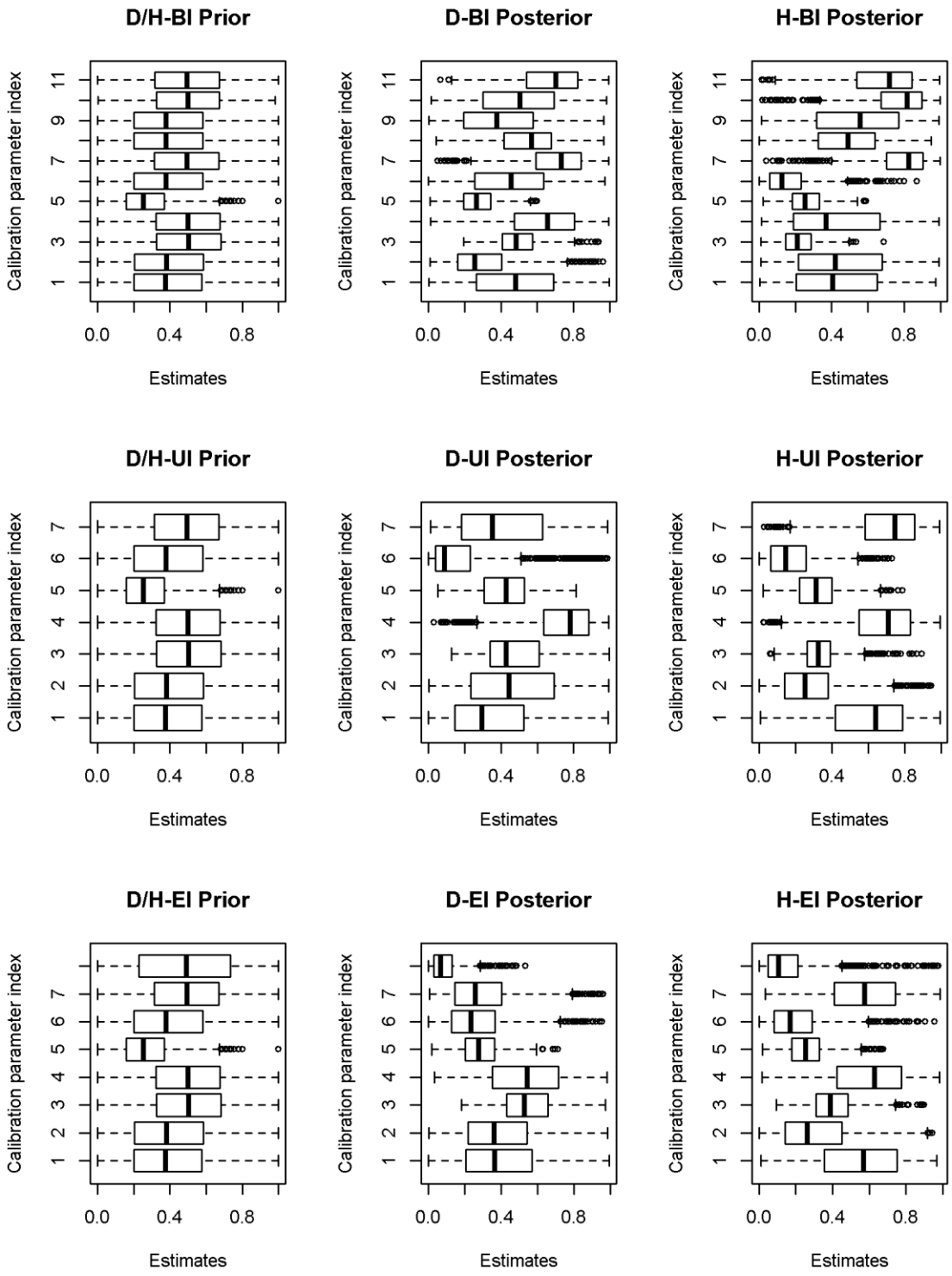


Figure 4.8 Prior and posterior estimates of calibration parameters

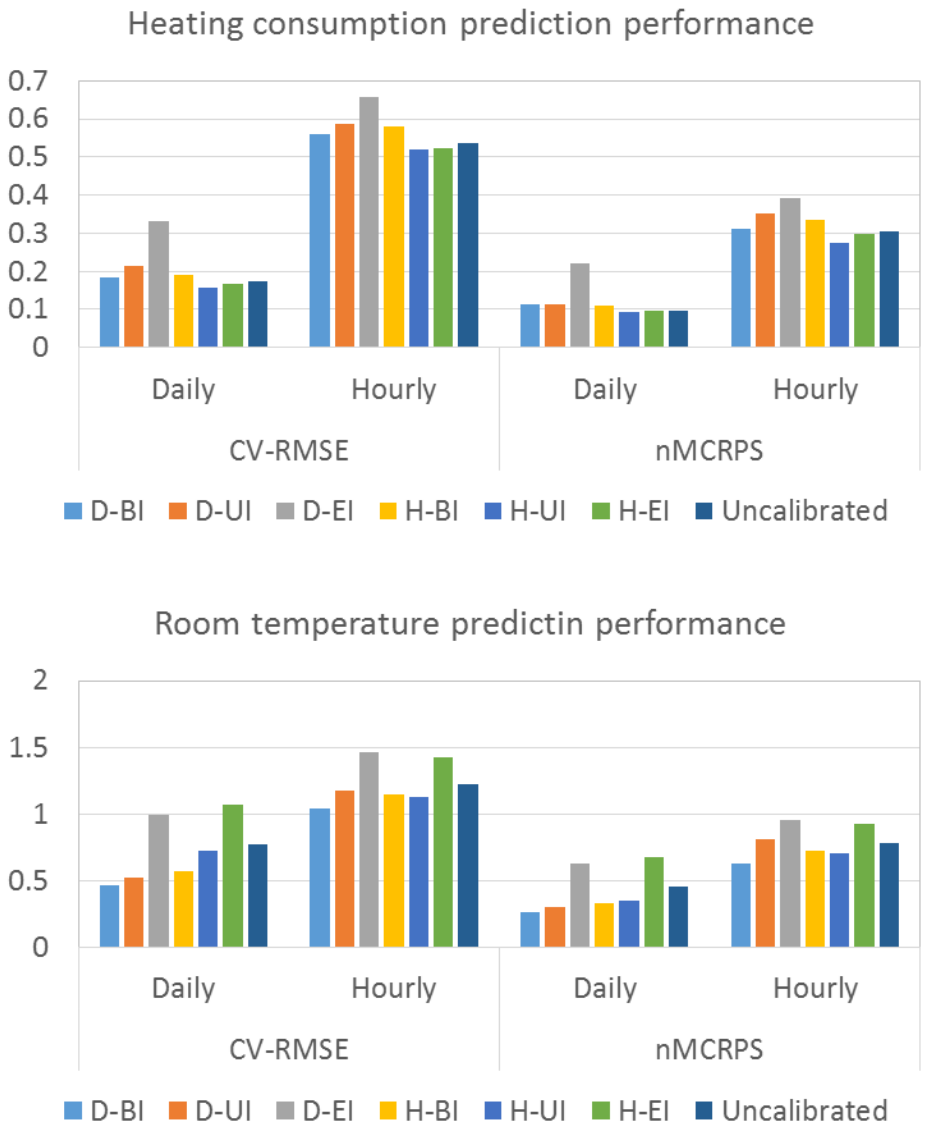


Figure 4.9 Result of physical model prediction assessment

4.4.2 Hypothetical intervention analysis and decision risk assessment

A hypothetical intervention analysis is designed to test the framework with respect to evaluating the validity of the calibrated models to support risk-conscious building performance management practice. In contrast to the design of empirical validation experiment, this risk is associated with system underperformance of an intervention

decision. Therefore, the calibrated models take the role of the internal model in the first case study and represent the enclosed information set, and the intended application changes from validation of external models to evaluation of intervention performances.

A hypothetical intervention is proposed to reduce heating consumption while maintaining similar thermal comfort conditions, which reduces the VT loop hot water supply temperature by 5°C, unless after doing so the temperature would fall below 25°C, the actual minimum supply temperature. Simulation of the test week, 03/20/2017-03/26/2017, is performed by feeding the calibration parameter posteriors into the physical model and applying the interventions' respective adjustments. Actual weather, usage, and heating operation conditions are used. Therefore, the uncertainty only comes from parameter uncertainties and model bias. Parameter uncertainties come from either prior or posterior estimates depending on whether it is calibrated in each calibration scenario. Model bias is considered by adding a random permutation of errors from the testing period to the physical model prediction in each simulation instance; this is performed 100 times for each of the 100 posterior sample points and thus results in a probabilistic prediction of 10,000 points in total.

Two outcomes are considered in the analysis regarding energy use and discomfort respectively: the one-week total heating consumption, and the hourly average meeting room discomfort degree-hours per heating-hour. In the business-as-usual scenario, the heating consumption of the business-as-usual scenario is the actual observation, whereas the discomfort is calculated based on the observed meeting room temperature and estimated room temperature setpoint; in the intervention scenario both of them are estimated.

A risk measure is defined in this case study to assess model validity based on the predicted intervention underperformance probability. Two decision rules are considered accordingly. The first decision rule (D1) is that the intervention will be implemented unless the underperformance probability that either of the following two events happens exceeds 50%: 1) the projected heating consumption is larger than the observed heating consumption, 2) the projected discomfort degree-hours is greater than the estimated current discomfort degree-hours. The second decision rule (D2) has the same underperformance probability threshold, but allows a slightly compromised thermal comfort to reduce energy use, so the two events become: 1) the projected one-week heating consumption is larger than *90%* of the observed heating consumption, 2) the projected discomfort degree-hours is greater than the estimated current discomfort degree-hours *plus 0.5°C-h*. The decision risk is defined as the probability of undesired outcomes, which is either the probability that the intervention underperforms when it is implemented, or the probability that the intervention is effective but not implemented. Given the underperformance probability p, the decision risk becomes:

$$DecisionRisk(D1/2) = \begin{cases} p, & p \leq 0.5 \\ 1-p, & p > 0.5 \end{cases} \tag{23}$$

The underperformance probability of each output, and corresponding decision risks of both decision rules from the six calibration scenarios are shown in Figure 4.10. The results suggest that the relevance of calibration, equivalently the validity of a calibrated model to support a decision-making, depends not only on the physical process, but also on the specific decision rules, the perceived value of each outcome, and the risk tolerance of the decision maker. In this intervention analysis, all the models can be deemed valid if

discomfort is the primary concern and cannot be compromised, as the risks of the first decision rule is universally negligible. In contrast, in the case of the second decision rule, the validity of model depends also on the sensitivity of the perceived value of outcome to the revealed risks as well as their differences. If taking a 10% decision risk as a threshold, it is concluded that only the models calibrated in scenario D-BI and D-UI appear to be valid to support the decision. In summary, the proposed explicit calculation of decision risk by sampling on model parameter uncertainties and prediction errors proves to be an effective way to represent model validity and data informativeness for risk-conscious BPS applications, which tentatively accepts the second research hypothesis.

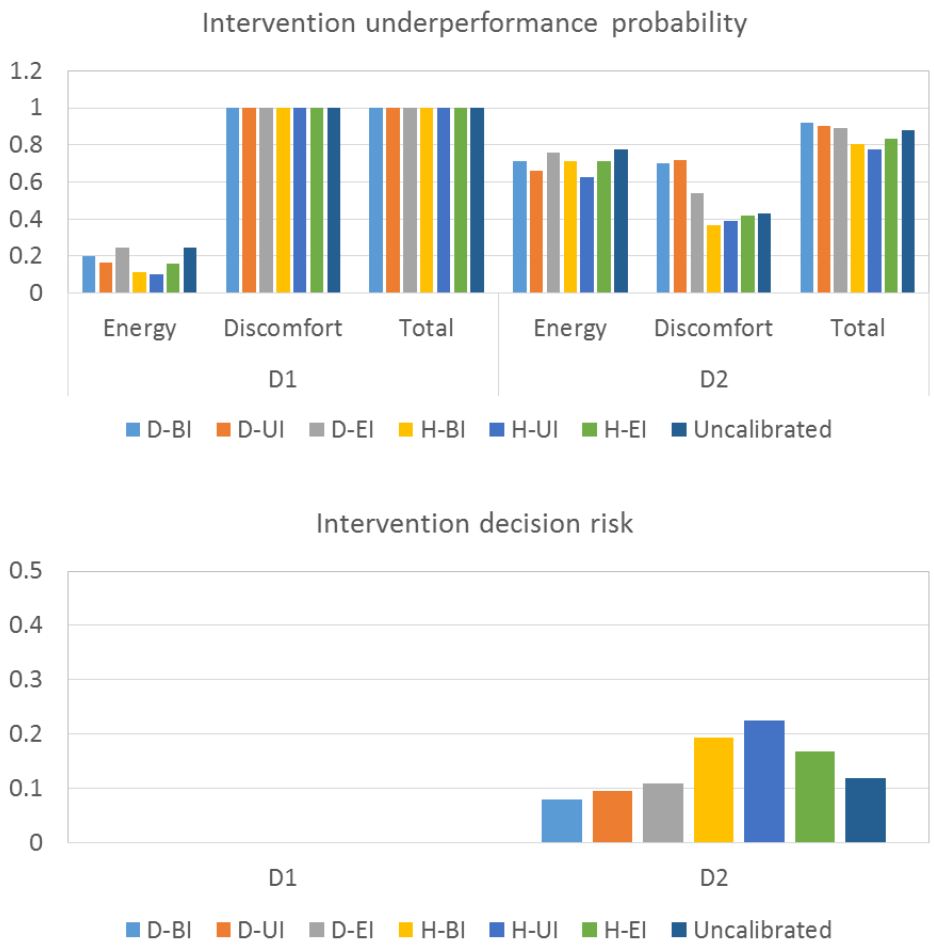


Figure 4.10 Decision risks regarding energy and discomfort outcomes

In addition, as the model uncertainties do not include those associated with weather, usage, and system operations because of the use of actual conditions, it is expected that the underperformance probability will move toward 0 or 1, and thus the misjudgement risks on individual outcomes will decrease, when a model's extrapolating accuracy toward future observations is increasing. This is because a perfect model that fits exactly on the observations and has no uncertainties would provide a zero decision risk associated with model accuracy. In this sense, the linear correlation between the risks of each decision rule regarding the individual outcomes, and model accuracy with respect to the testing dataset under either the CV-RMSE or the CRPS of all the six calibration scenarios, can serve as a

crude assessment of whether latter truly reflect the model's extrapolating accuracy. The calculated linear correlation coefficients between each group of accuracy metric values and each type of individual decision risk are shown in Figure 4.11. While a value close to 1 means that the accuracy metrics truly reflect the extrapolating accuracy at least in an ordinal manner, the results suggest that none of these accuracy metrics appears to be consistently indicative of the extrapolating accuracy. This observation suggests the potential inadequacy of the information and data particular as they only reflect the current condition, and the extrapolation risk in intervention analysis when the physical model is calibrated primarily through inverse modelling. Model validity largely depends on the specific decision-making scenarios, can benefit from explicit risk assessment of individual decisions through Bayesian inference, and more importantly is not decisively determined by model accuracy. Nevertheless, it is expected that forward uncertainty quantification with detailed monitoring of building sub-systems is a more robust alternative in constraining model uncertainties and improving extrapolation prediction accuracy, which is vital for a universally reliable physical model to support BPS applications in real practice.

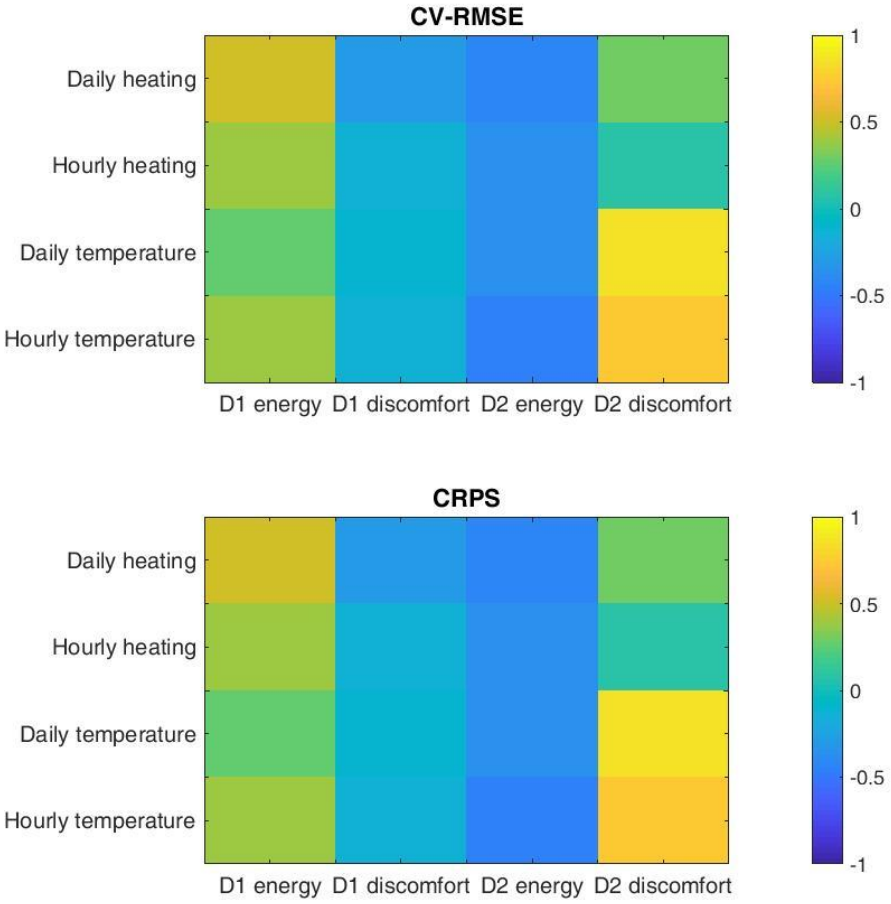


Figure 4.11 Linear correlation between accuracy metrics and decision risk

CHAPTER 5. CONCLUSIONS AND FUTURE WORK

5.1 Summary and conclusions

This dissertation addresses data informativeness in building performance simulation (BPS) applications by proposing a framework that builds upon uncertainty propagation using detailed measurements, inverse modelling using Bayesian inference, variants of the continuous rank probability score (CRPS) as the probabilistic accuracy metrics, and explicit risk assessment through sampling on model parameter uncertainties and prediction errors. Two case studies were provided to demonstrate the framework, which include the design of empirical validation experiments and a hypothetical intervention analysis of a campus building with hydronic heating. The results demonstrate the effectiveness of forward and inverse uncertainty quantification in improving model predictions, and the effectiveness of explicitly risk assessment in validating models and representing data informativeness for specific applications. The results also show that the CRPS is a more informative accuracy metric than traditional counterparts especially when this metric is used for risks associated with existing observations. Conversely, when using inverse modelling to make inferences about future observations with potential extrapolations, model accuracy toward existing observations may not necessarily reflect the underlying risks and reveal the inadequacy of the information and data. In summary, this dissertation has demonstrated that:

I. The proposed calibration methods make improved use of data in constraining uncertainty and improving prediction.

II. The proposed explicit risk assessment improves the representation of model validity and data informativeness under uncertainty.

5.2 Recommendations for future study

Data informativeness is and will continue to be an important area of study for building performance simulation applications in future real practice. This dissertation makes a limited attempt to systematically address this issue because of time and resource constraints, and the ultimate goal of the research is to establish a complete framework to understand, represent, and assess data informativeness in building performance simulation applications. Some immediate future work would include:

5.2.1 Empirical validation experiment design

An immediate future work is to test the proposed empirical validation methodology with real measurement data and additional validation experiments. It would also interesting to apply Bayesian calibration to the baseline model and compare the result with that of forward uncertainty propagation for an improved understanding of calibration verification. As being mentioned previously, estimation of the true validation risks through extensive investigations on a variety of model form errors, and potentially systematic quantification of modeler's bias via human-subject studies would also be important to improve the empirical validation methodology.

5.2.2 Hydronic heating system intervention analysis

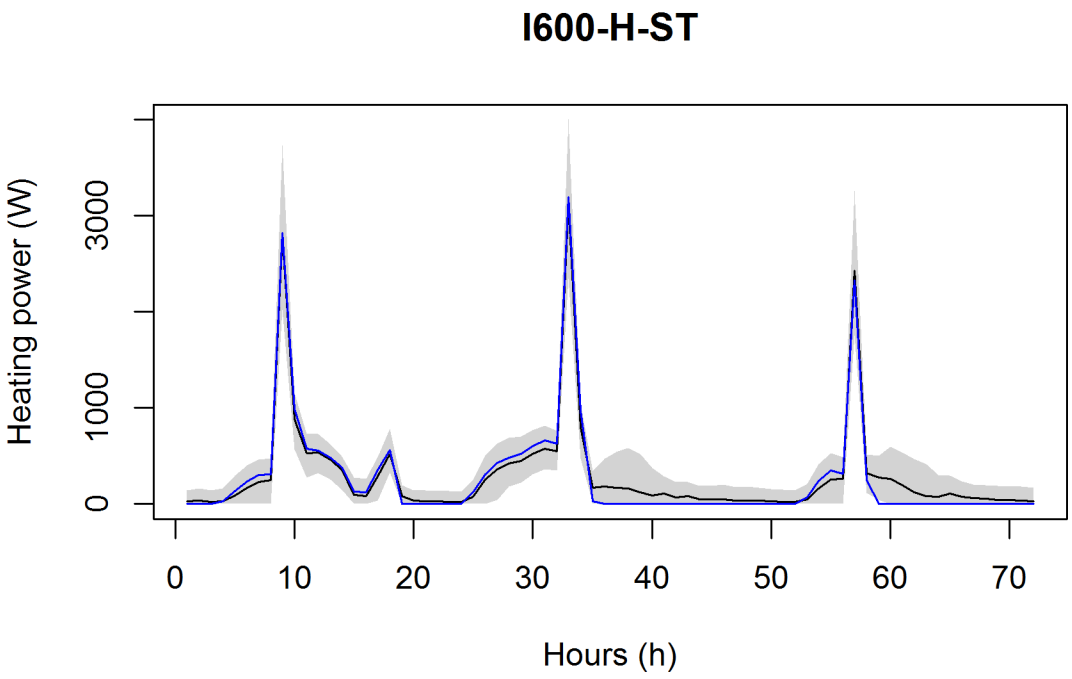
It is recommended to explore the use of explicit model form uncertainty quantification to inform Bayesian calibration through improved priors of model bias, and

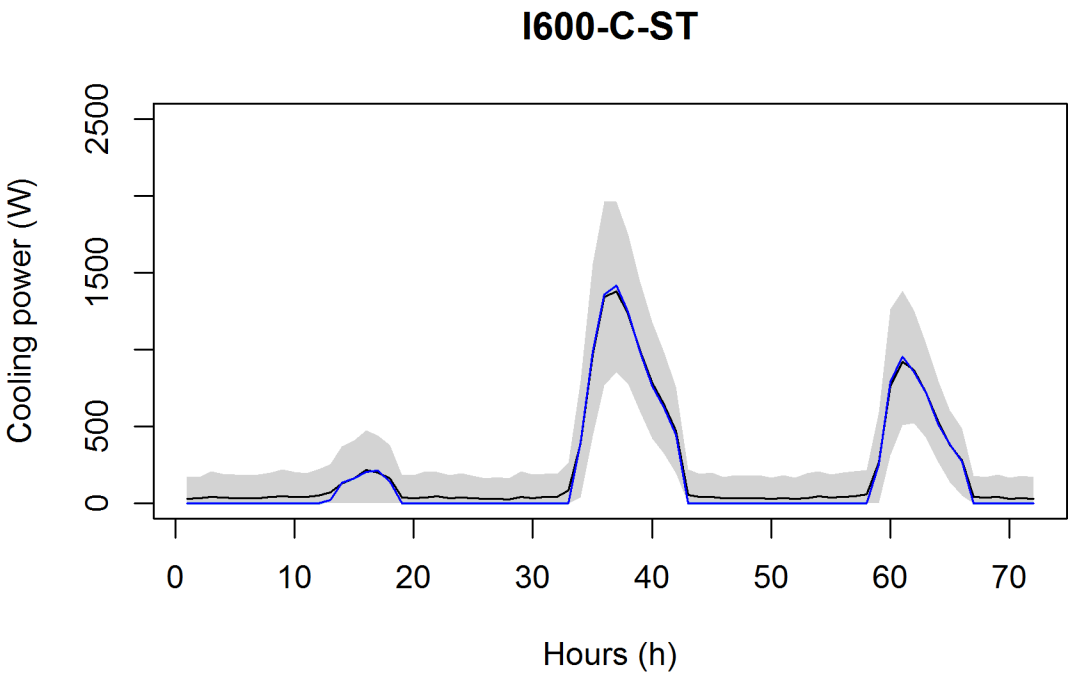
test data informativeness with increased sample size especially on daily monitoring data. An efficient way to combine data at different time resolutions, for example a hourly calibration on the mean profiles and a daily calibration to account for day-to-day variations would also have great potential in real practice. Investigation on appropriate emulators of the physical model for hourly data would be another important area of research. Finally, an explorative study on Bayesian calibration using synthetic measurements would be appealing to test and validate this inverse modelling method.

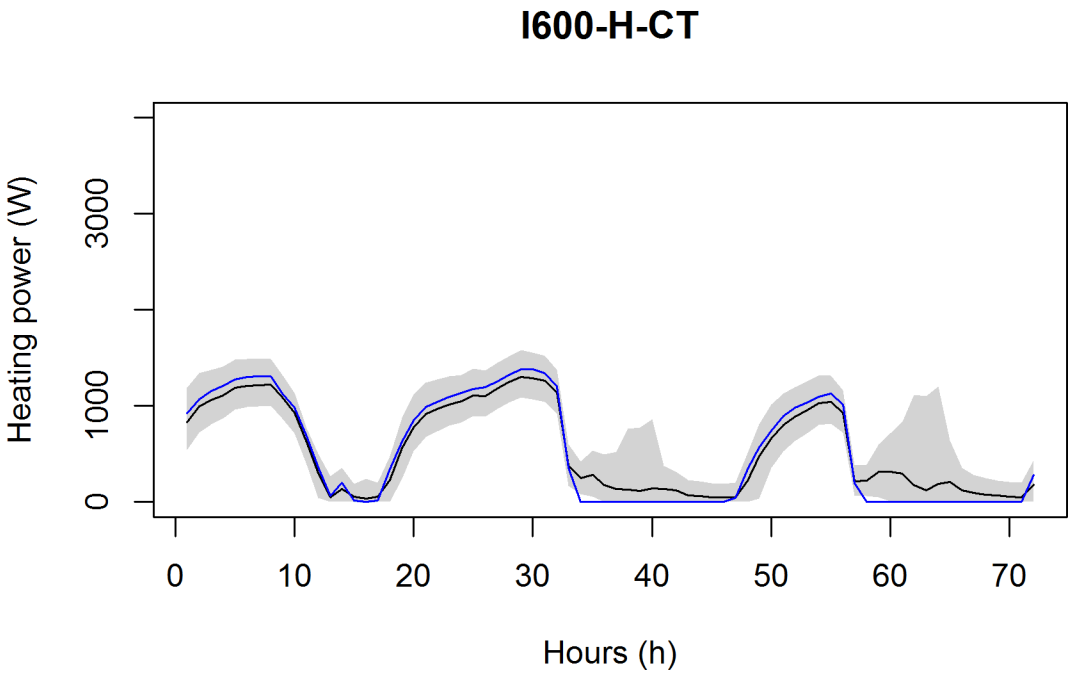
5.2.3 The general model calibration and validation framework

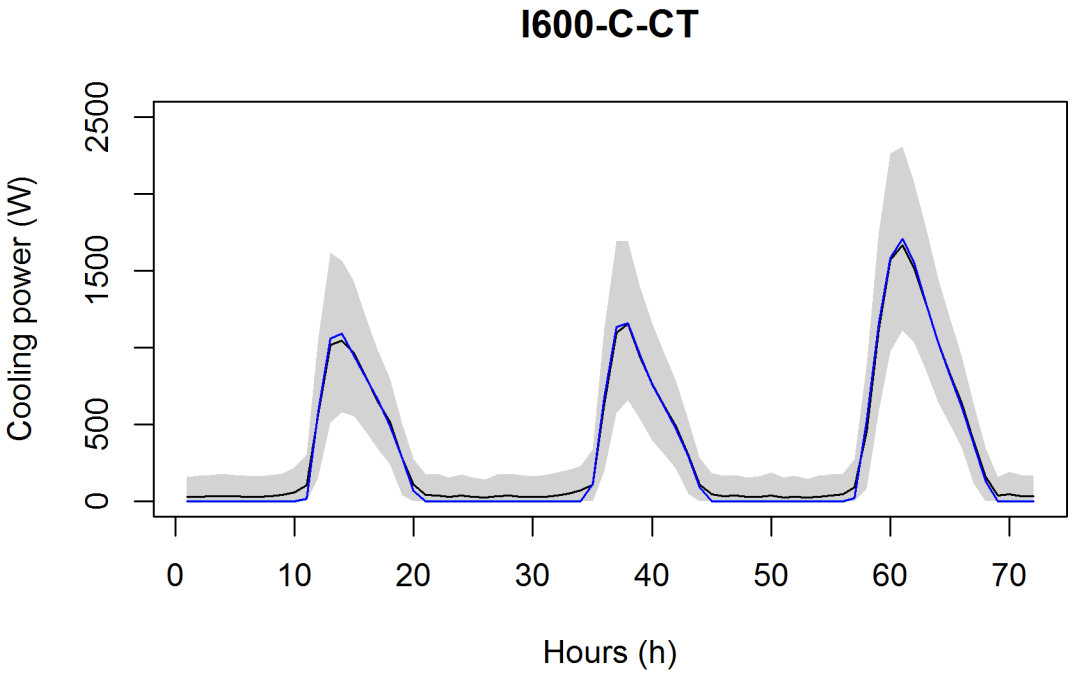
The most recommended future work is to implement and test the new Bayesian calibration framework from Tuo and Wu (2016) with respect to its efficiency and effectiveness in real building performance simulation applications. In the meantime, a framework to efficiently combine forward and inverse modelling methods from a Bayesian perspective is worth further study. Finally, practical insights can be obtained by applying the proposed framework to a more realistic practice where data informativeness directly relates to decision-making processes and interests of stakeholders.

APPENDIX A. DETAILED TABLES AND FIGURES






I600-H-CT
Heating power (W)
0
1000
3000
0
10
20
30
40
50
60
70
Hours (h)


I600-C-CT
Cooling power (W)
0
500
1500
2500
0
10
20
30
40
50
60
70
Hours (h)

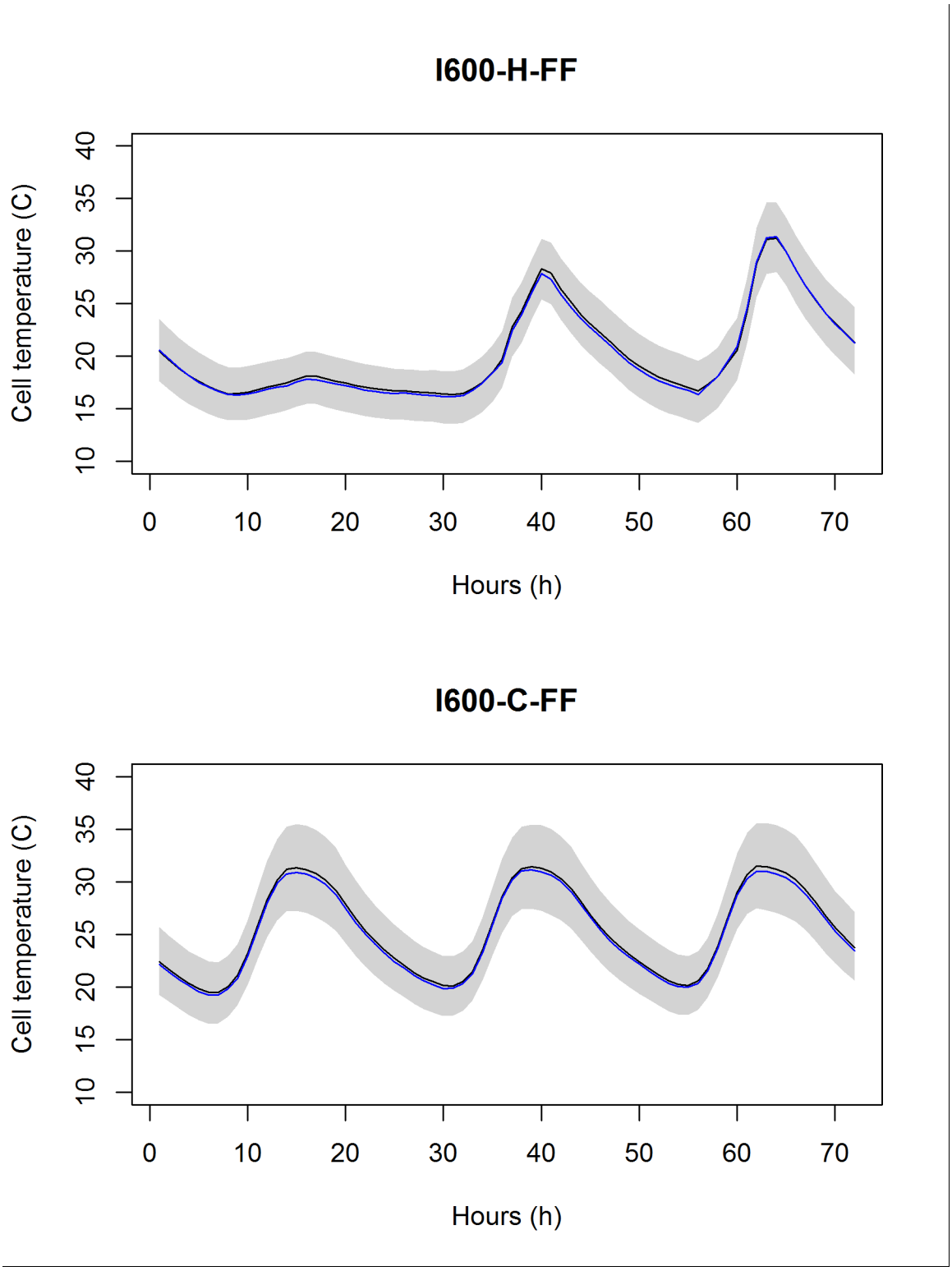
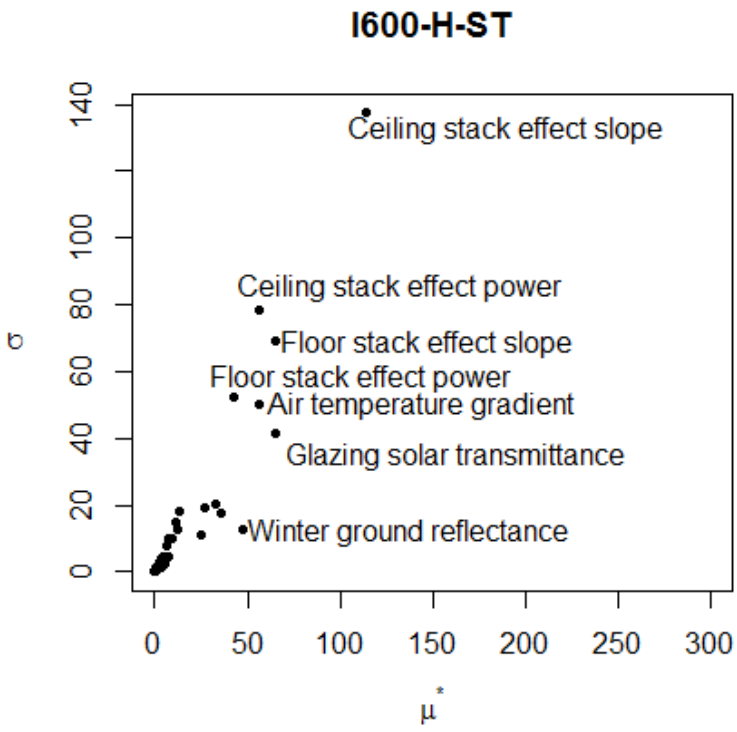
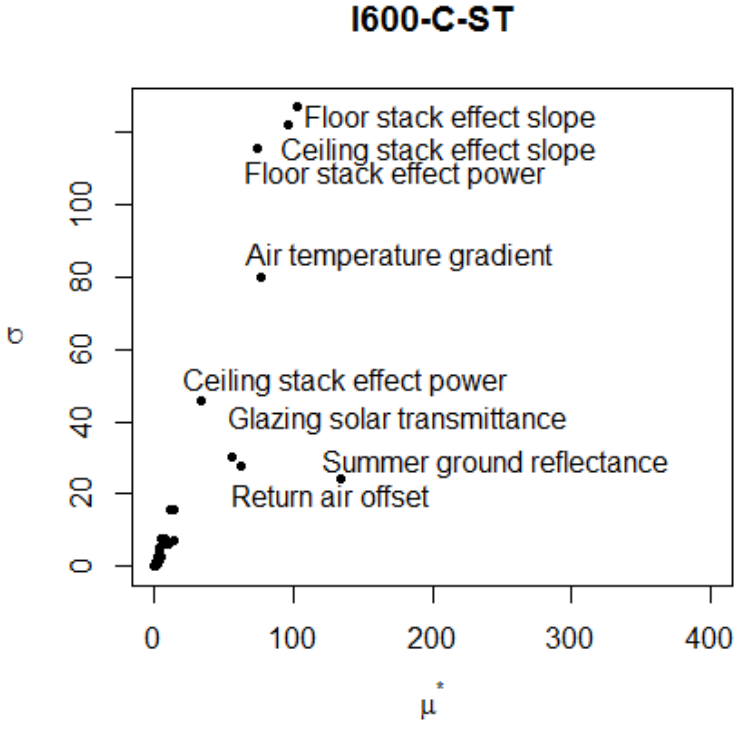
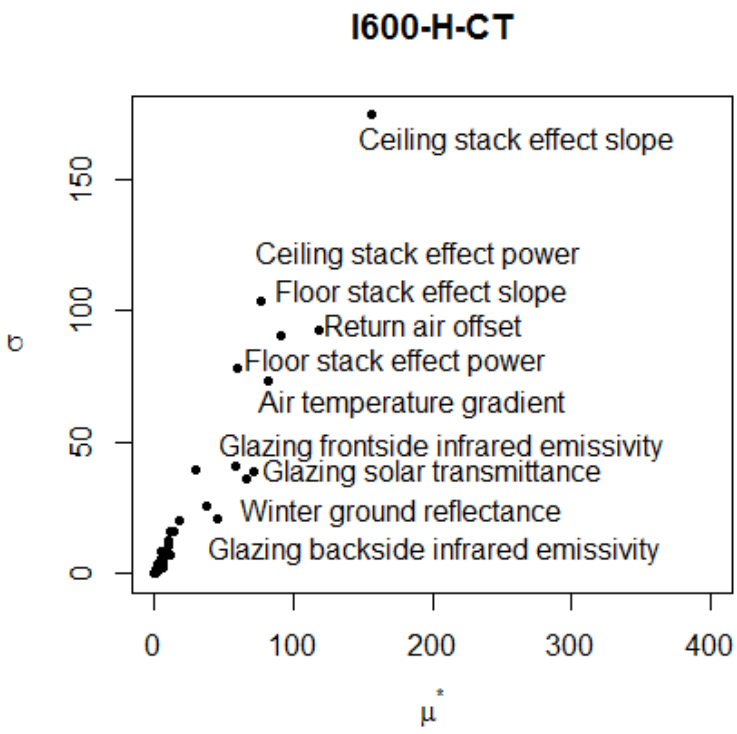
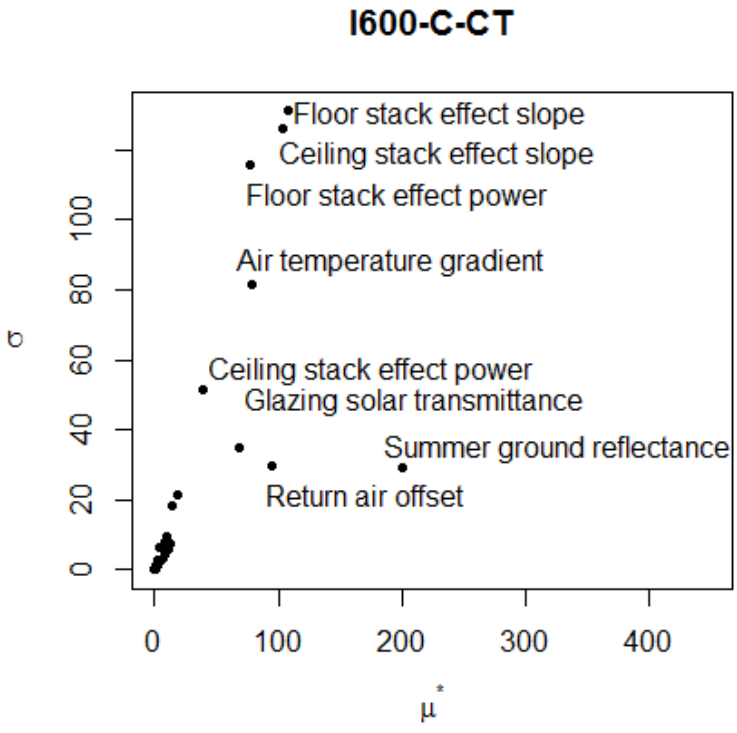


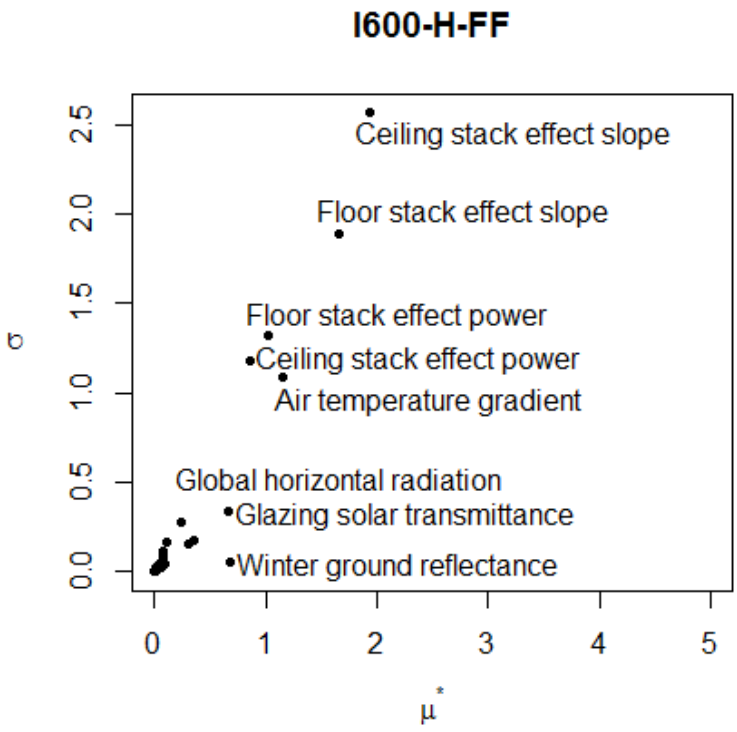
Figure A.1 Result of baseline internal model with 95% confidence interval


I600-H-ST
Ceiling stack effect slope
Ceiling stack effect power
Floor stack effect slope
Floor stack effect power
Air temperature gradient
Glazing solar transmittance
Winter ground reflectance
σ
μ*
0
20
40
60
80
100
140
0
50
100
150
200
250
300


I600-C-ST
Floor stack effect slope
Ceiling stack effect slope
Floor stack effect power
Air temperature gradient
Ceiling stack effect power
Glazing solar transmittance
Summer ground reflectance
Return air offset
σ
μ*
0
20
40
60
80
100
0
100
200
300
400


I600-H-CT
Ceiling stack effect slope
Ceiling stack effect power
Floor stack effect slope
Return air offset
Floor stack effect power
Air temperature gradient
Glazing frontside infrared emissivity
Glazing solar transmittance
Winter ground reflectance
Glazing backside infrared emissivity
σ
0
50
100
150
0
100
200
300
400
μ*


I600-C-CT
Floor stack effect slope
Ceiling stack effect slope
Floor stack effect power
Air temperature gradient
Ceiling stack effect power
Glazing solar transmittance
Summer ground reflectance
Return air offset
σ
0
20
40
60
80
100
0
100
200
300
400
μ*



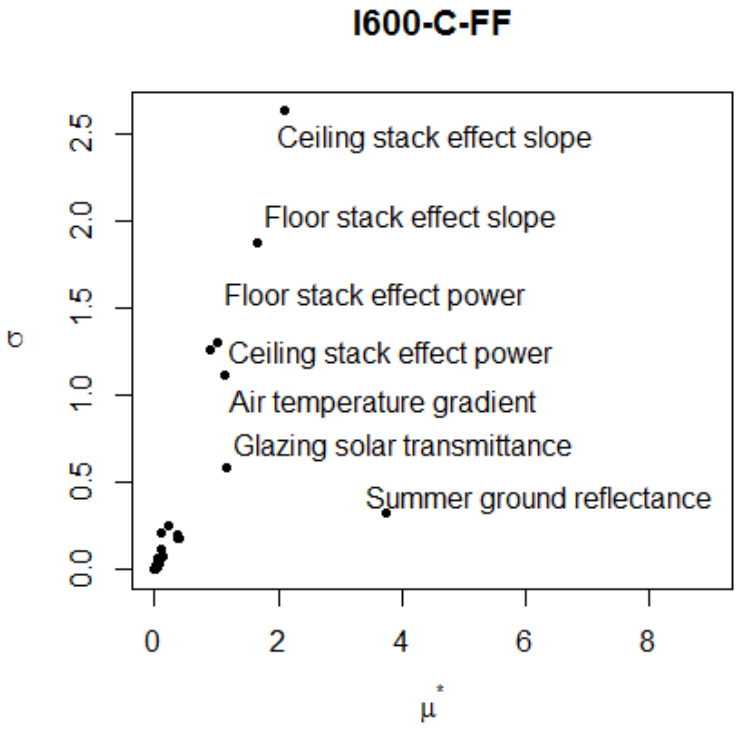


Figure A.2 Result of sensitivity analysis of baseline internal model

Table A.1 I600-H-ST sensitivity analysis result of the top 20 parameters

Parameter	Main effects	Interactions
Ceiling stack effect slope	1.14e+02	1.37e+02
Glazing solar transmittance	6.49e+01	4.16e+01
Floor stack effect slope	6.47e+01	6.93e+01
Air temperature gradient	5.66e+01	5.02e+01
Ceiling stack effect power	5.64e+01	7.84e+01
Winter ground reflectance	4.72e+01	1.30e+01
Floor stack effect power	4.22e+01	5.21e+01
Return air offset	3.58e+01	1.74e+01
Glazing front side infrared emissivity	3.31e+01	2.03e+01
Polyiso 4.5” conductivity	2.68e+01	1.95e+01
Glazing back side infrared emissivity	2.51e+01	1.10e+01
Acous tile 3/4” specific heat	1.31e+01	1.80e+01
Global horizontal radiation	1.18e+01	1.26e+01
Plywood 1/2” with steel sheet specific heat	1.14e+01	1.49e+01
Wall stack effect slope	9.18e+00	1.01e+01
Plywood 1/2” with steel sheet solar absorptance	7.63e+00	4.68e+00
Wind effect constant	7.21e+00	9.96e+00
Gypsum board specific heat	7.19e+00	9.55e+00
Wall stack effect power	6.60e+00	8.12e+00
Glazing front side solar reflectance	4.94e+00	2.37e+00

Table A.2 I600-H-CT sensitivity analysis result of the top 20 parameters

Parameter	Main effects	Interactions
Ceiling stack effect slope	1.56e+02	1.75e+02
Return air offset	1.18e+02	9.29e+01
Floor stack effect slope	9.03e+01	9.08e+01
Air temperature gradient	8.11e+01	7.32e+01
Ceiling stack effect power	7.61e+01	1.04e+02
Glazing solar transmittance	7.15e+01	3.90e+01
Winter ground reflectance	6.58e+01	3.60e+01
Floor stack effect power	5.98e+01	7.84e+01
Glazing fron tside infrared emissivity	5.82e+01	4.10e+01
Glazing back side infrared emissivity	4.44e+01	2.07e+01
Polyiso 4.5” conductivity	3.68e+01	2.60e+01
Plywood 1/2” with steel sheet specific heat	2.93e+01	3.97e+01
Acous tile 3/4” specific heat	1.73e+01	2.04e+01
Wall stack effect slope	1.41e+01	1.64e+01
Gypsum board specific heat	1.11e+01	1.59e+01
Plywood 1/2” with steel sheet solar absorptance	1.09e+01	7.39e+00
Wind effect constant	9.91e+00	1.30e+01
Wall stack effect power	9.58e+00	1.10e+01
Global horizontal radiation	9.29e+00	1.16e+01
Gypsum board solar absorptance	7.38e+00	6.18e+00

Table A.3 I600-H-FF sensitivity analysis result of the top 20 parameters

Parameter	Main effects	Interactions
Ceiling stack effect slope	1.94e+00	2.56e+00
Floor stack effect slope	1.66e+00	1.89e+00
Air temperature gradient	1.16e+00	1.09e+00
Floor stack effect power	1.02e+00	1.32e+00
Ceiling stack effect power	8.51e-01	1.18e+00
Winter ground reflectance	6.78e-01	5.35e-02
Glazing solar transmittance	6.60e-01	3.38e-01
Glazing back side infrared emissivity	3.48e-01	1.77e-01
Glazing front side infrared emissivity	3.07e-01	1.60e-01
Global horizontal radiation	2.41e-01	2.76e-01
Wall stack effect slope	1.01e-01	1.69e-01
Plywood 1/2” with steel sheet solar absorptance	8.73e-02	5.02e-02
Wall stack effect power	8.14e-02	1.14e-01
Wind effect constant	7.80e-02	9.46e-02
Horizontal diffuse radiation	6.80e-02	6.78e-02
Glazing front side solar reflectance	5.07e-02	2.59e-02
Gypsum board solar absorptance	4.23e-02	3.52e-02
South wall insulation layer conductivity	3.89e-02	2.22e-02
Wind effect slope	3.63e-02	3.17e-02
Polyiso 4.5” conductivity	3.48e-02	4.35e-02

Table A.4 I600-C-ST sensitivity analysis result of the top 20 parameters

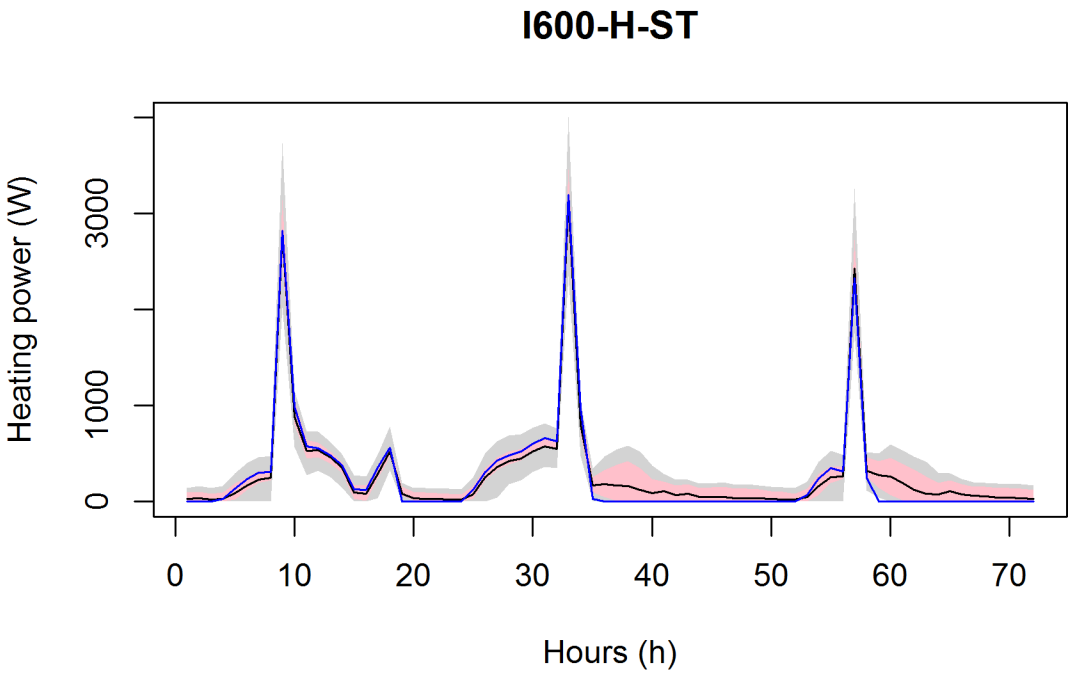
Parameter	Main effects	Interactions
Summer ground reflectance	2.00e+02	2.95e+01
Floor stack effect slope	1.08e+02	1.31e+02
Ceiling stack effect slope	1.03e+02	1.26e+02
Return air offset	9.44e+01	2.96e+01
Air temperature gradient	7.90e+01	8.16e+01
Floor stack effect power	7.78e+01	1.16e+02
Glazing solar transmittance	6.80e+01	3.48e+01
Ceiling stack effect power	3.87e+01	5.17e+01
Wall stack effect slope	1.78e+01	2.15e+01
Wall stack effect power	1.40e+01	1.83e+01
Glazing front side infrared emissivity	1.29e+01	7.66e+00
Glazing back side infrared emissivity	1.07e+01	6.15e+00
Wind effect constant	9.50e+00	9.79e+00
Polyiso 4.5” conductivity	8.61e+00	4.40e+00
Plywood 1/2” with steel sheet specific heat	7.50e+00	7.98e+00
Plywood 1/2” with steel sheet solar absorptance	6.16e+00	3.30e+00
Global horizontal radiation	5.95e+00	6.65e+00
Acous tile 3/4” specific heat	4.97e+00	6.23e+00
Glazing front side solar reflectance	4.14e+00	2.26e+00
Gypsum board specific heat	4.05e+00	6.36e+00

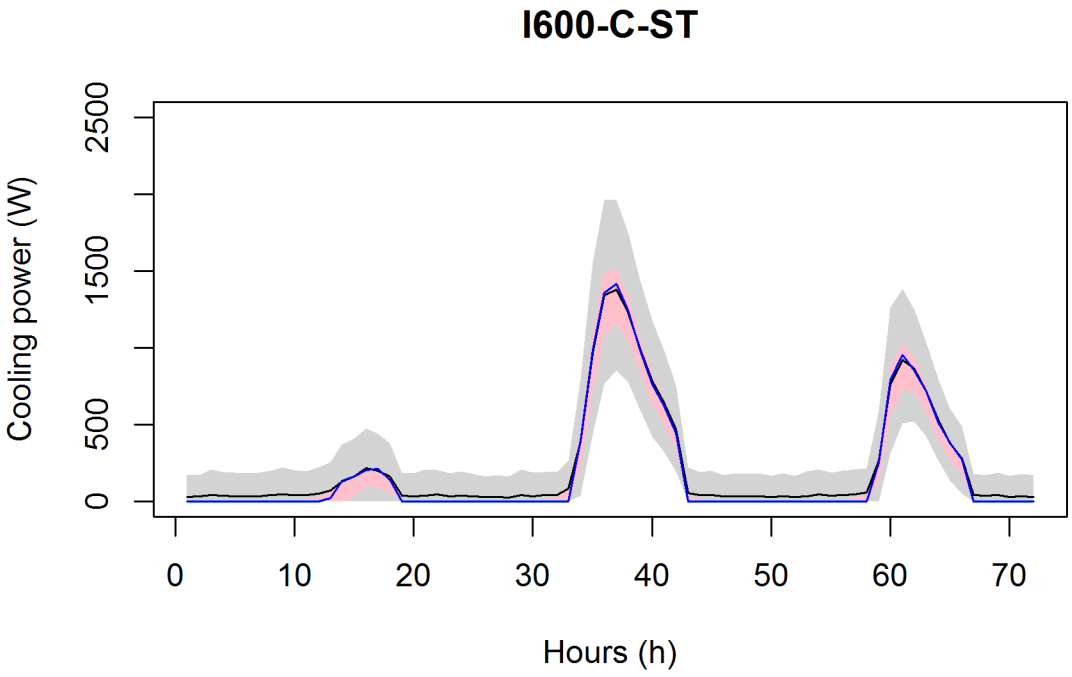
Table A.5 I600-C-CT sensitivity analysis result of the top 20 parameters

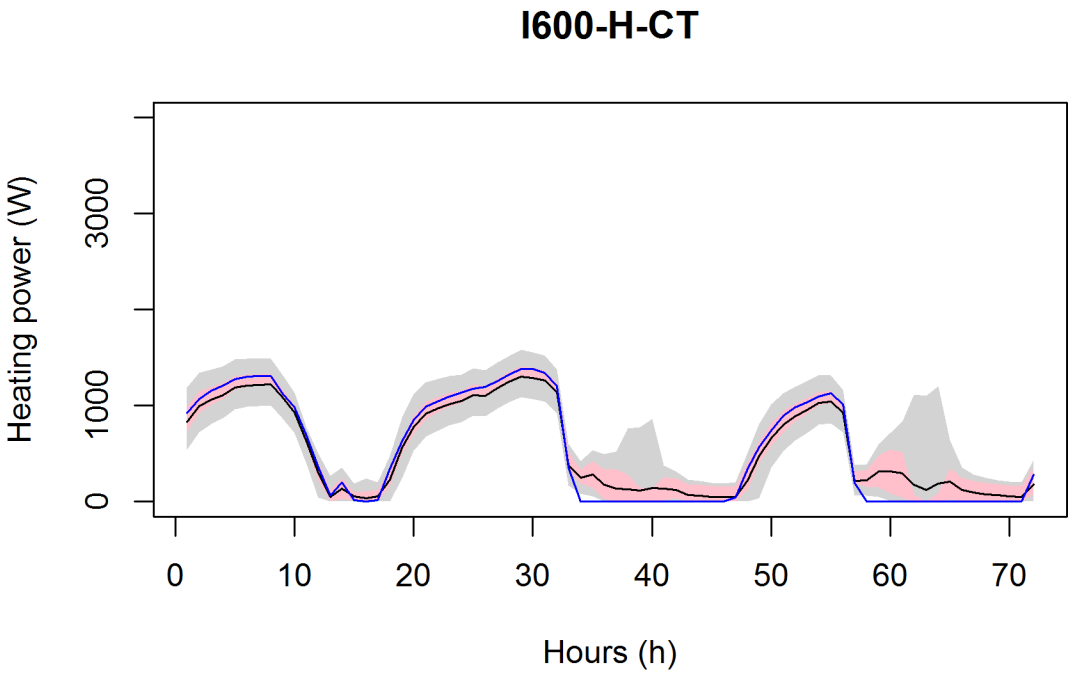
Parameter	Main effects	Interactions
Summer ground reflectance	1.34e+02	2.44e+01
Floor stack effect slope	1.03e+02	1.27e+02
Ceiling stack effect slope	9.56e+01	1.22e+02
Air temperature gradient	7.69e+01	8.00e+01
Floor stack effect power	7.33e+01	1.15e+02
Return air offset	6.19e+01	2.75e+01
Glazing solar transmittance	5.55e+01	3.01e+01
Ceiling stack effect power	3.32e+01	4.57e+01
Glazing backside infrared emissivity	1.42e+01	7.48e+00
Wall stack effect slope	1.33e+01	1.58e+01
Wall stack effect power	1.17e+01	1.59e+01
Glazing front side infrared emissivity	1.02e+01	6.39e+00
Wind effect constant	7.69e+00	7.82e+00
Plywood 1/2” with steel sheet specific heat	5.84e+00	6.29e+00
Global horizontal radiation	4.90e+00	7.89e+00
Plywood 1/2” with steel sheet solar absorptance	4.55e+00	2.57e+00
Acous tile 3/4” specific heat	3.36e+00	4.35e+00
Gypsum board specific heat	3.21e+00	4.99e+00
Glazing front side solar reflectance	2.83e+00	1.56e+00
Wind effect slope	2.09e+00	2.71e+00

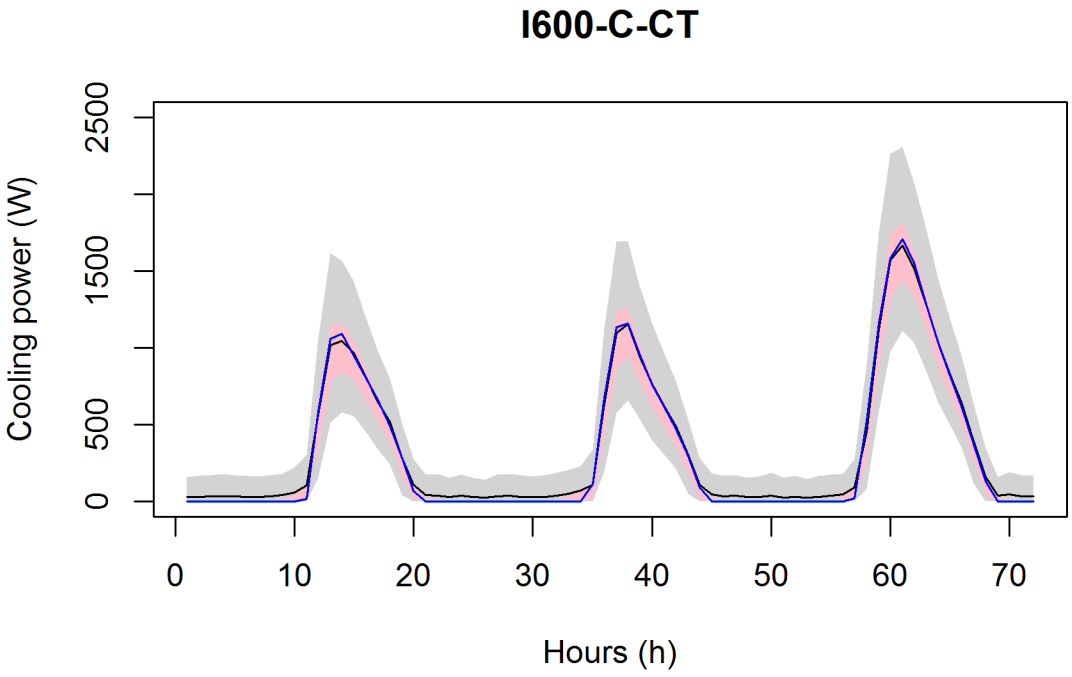
Table A.6 I600-C-FF sensitivity analysis result of the top 20 parameters

Parameter	Main effects	Interactions
Summer ground reflectance	3.73e+00	3.29e-01
Ceiling stack effect slope	2.10e+00	2.63e+00
Floor stack effect slope	1.65e+00	1.88e+00
Glazing solar transmittance	1.17e+00	5.86e-01
Air temperature gradient	1.13e+00	1.11e+00
Floor stack effect power	1.00e+00	1.30e+00
Ceiling stack effect power	8.84e-01	1.26e+00
Polyiso 4.5” conductivity	3.86e-01	1.81e-01
Glazing back side infrared emissivity	3.82e-01	1.78e-01
Glazing front side infrared emissivity	3.75e-01	2.02e-01
Wind effect constant	2.13e-01	2.51e-01
Plywood 1/2” with steel sheet solar absorptance	1.41e-01	7.78e-02
Wall stack effect slope	1.16e-01	2.15e-01
Wall stack effect power	9.13e-02	1.18e-01
Global horizontal radiation	8.17e-02	5.74e-02
Glazing front side solar reflectance	7.26e-02	3.79e-02
Wind effect slope	5.87e-02	6.89e-02
Structural insulation panel 7.25” conductivity	4.64e-02	3.37e-02
South wall insulation layer conductivity	4.57e-02	3.01e-02
Gypsum board solar absorptance	4.38e-02	3.27e-02


I600-H-ST
Heating power (W)
0
1000
3000
0
10
20
30
40
50
60
70
Hours (h)


I600-C-ST
Cooling power (W)
0
500
1500
2500
0
10
20
30
40
50
60
70
Hours (h)


I600-H-CT
Heating power (W)
0
1000
3000
0
10
20
30
40
50
60
70
Hours (h)


I600-C-CT
Cooling power (W)
0
500
1500
2500
0
10
20
30
40
50
60
70
Hours (h)

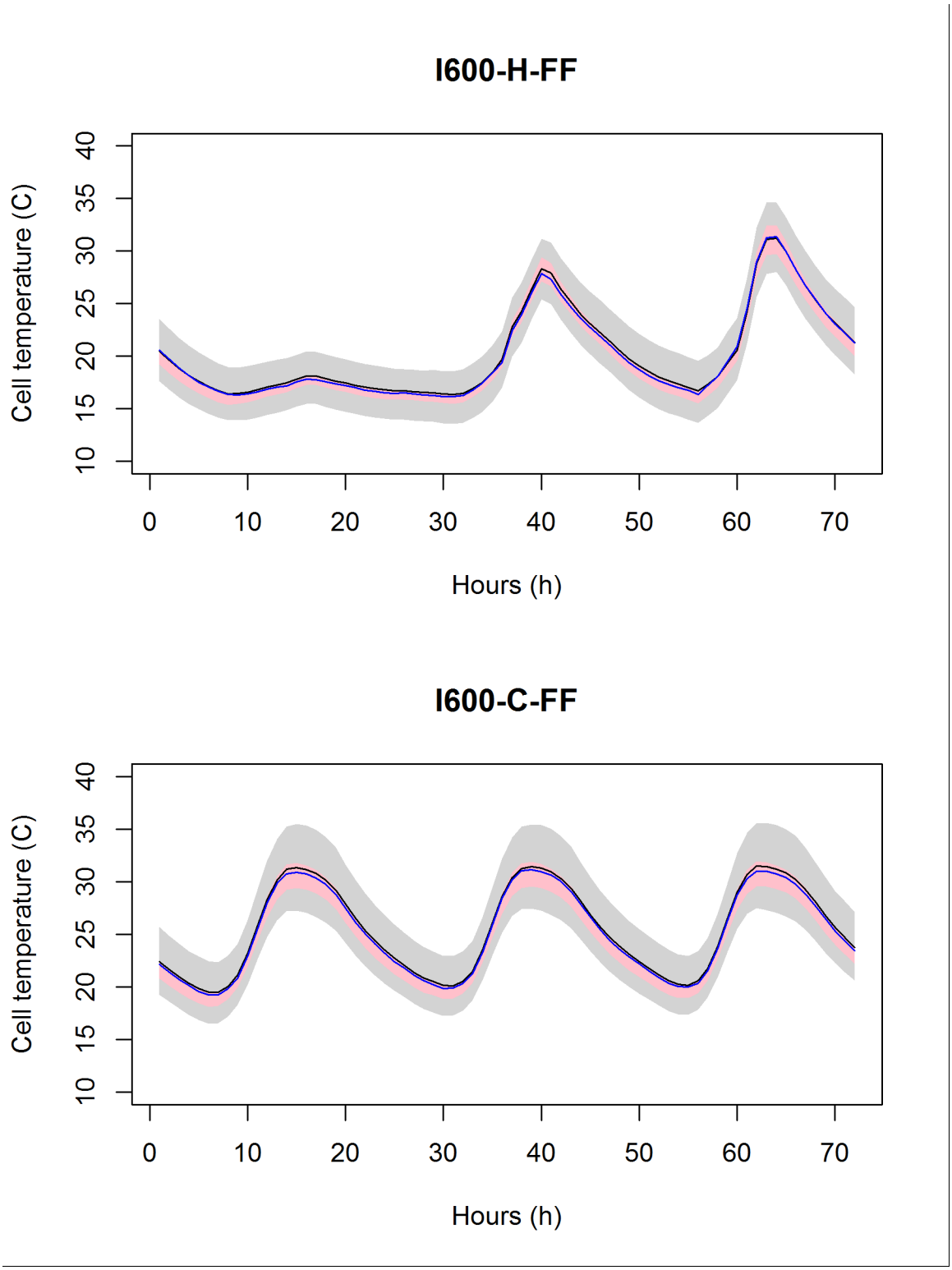


Figure A.3 Result of refined internal model with 95% confidence interval

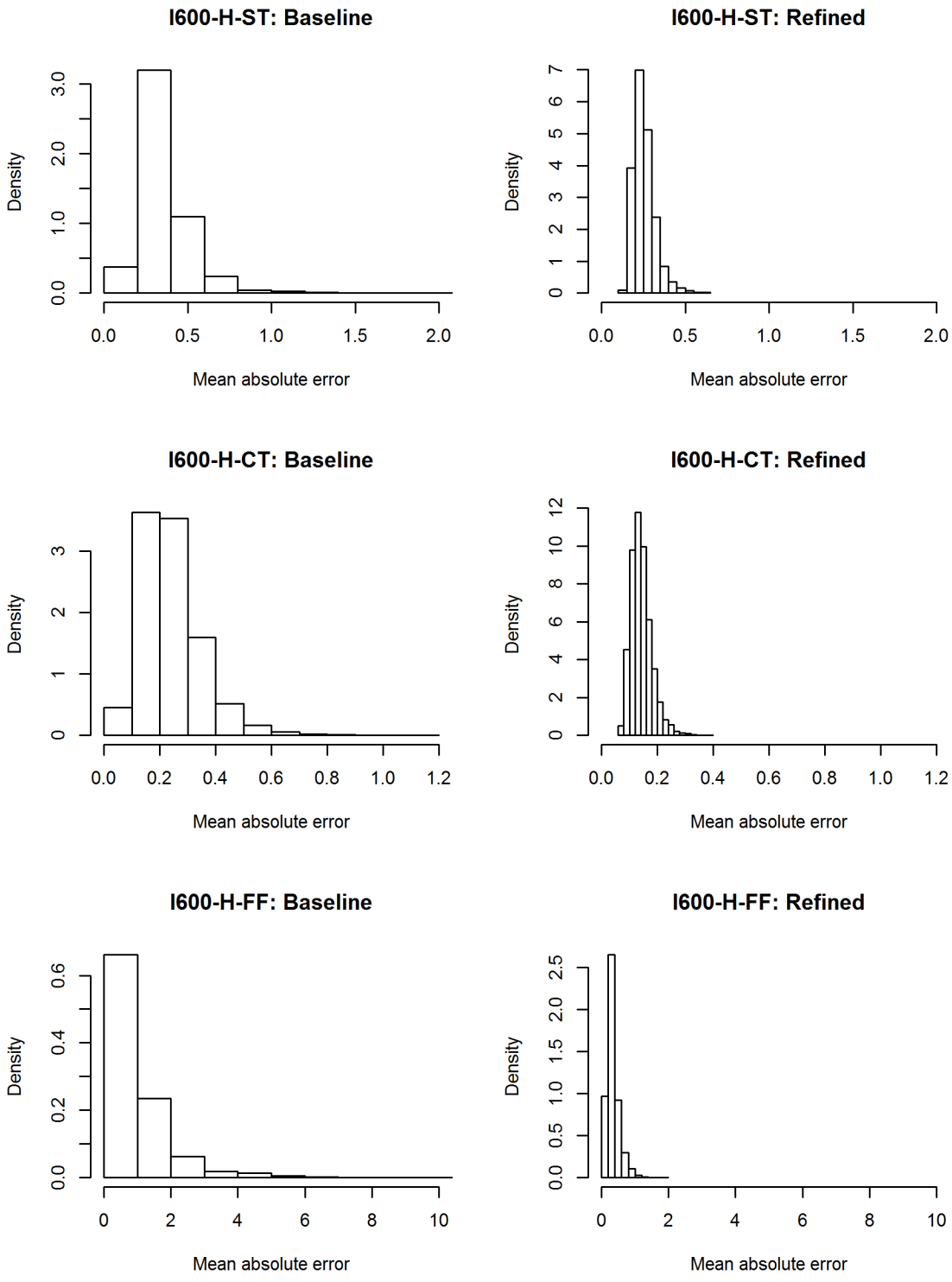


Figure A.4 Distribution of nMAE in AHU heating test

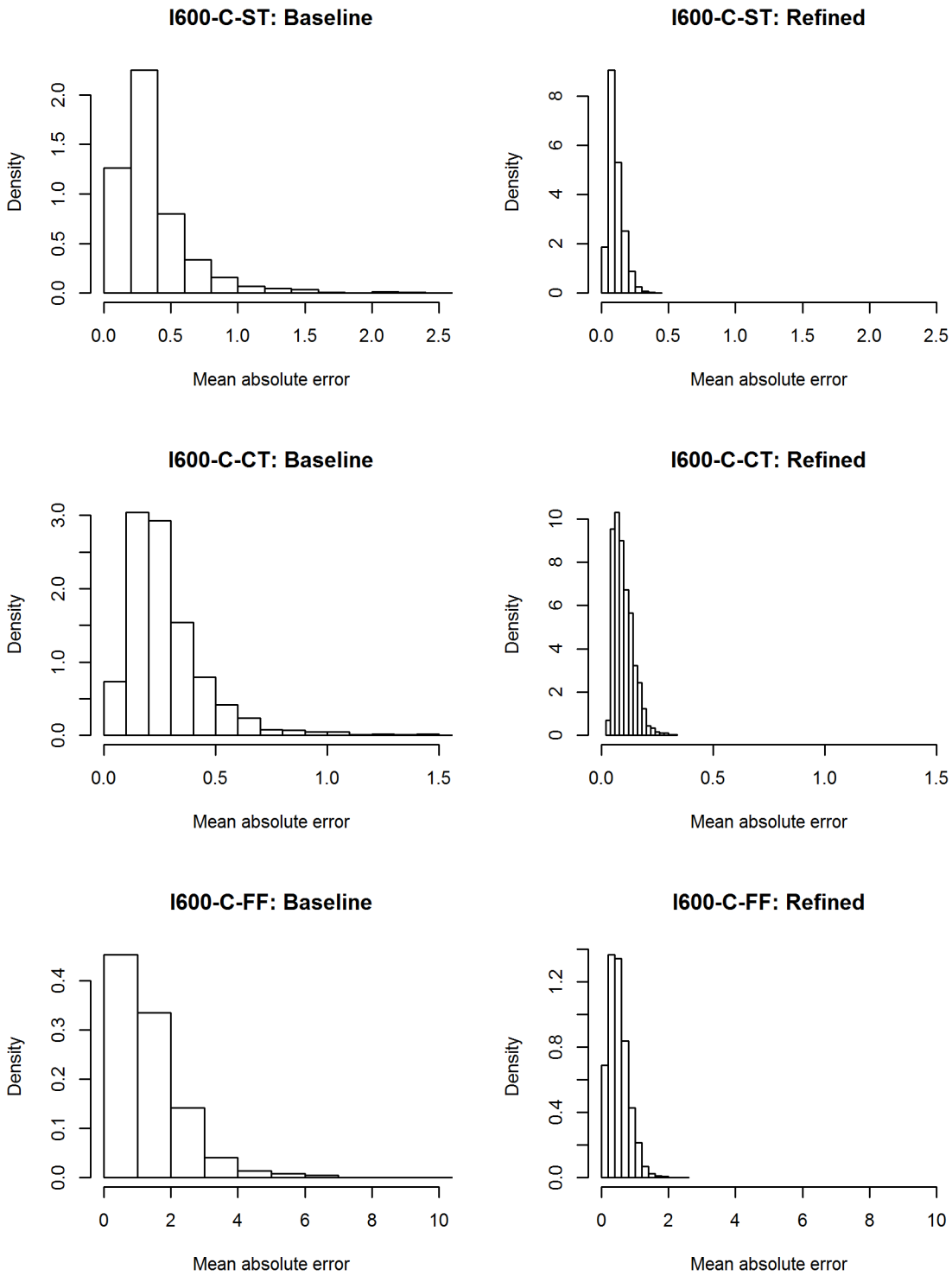


Figure A.5 Distribution of nMAE in AHU cooling test

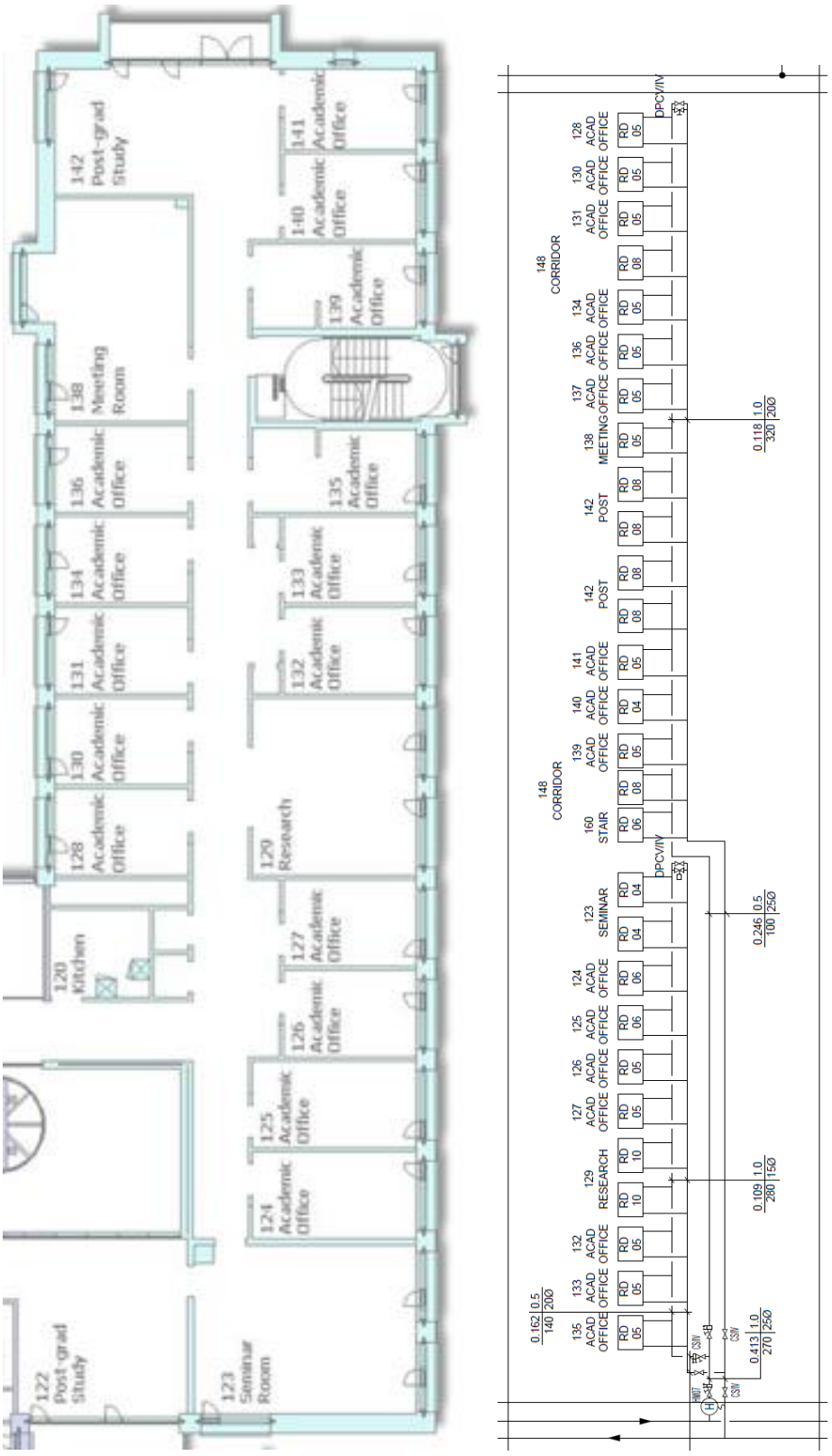
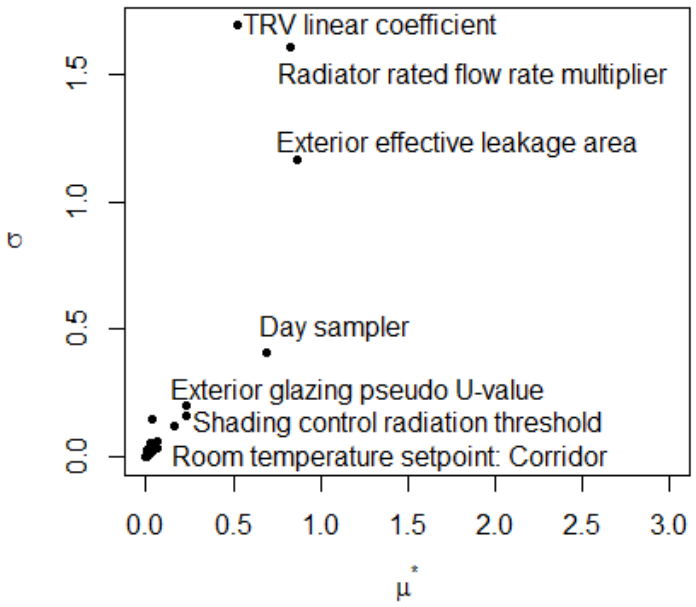


Figure A.6 Room plan and local heating loop schematics

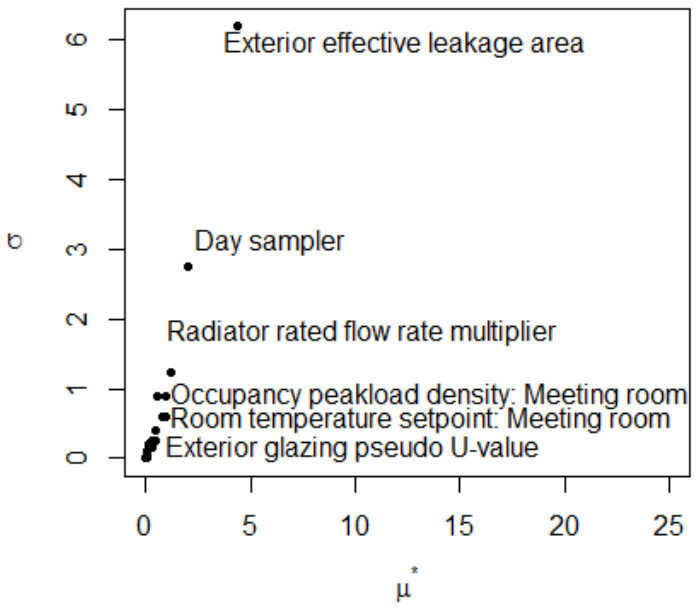
Table A.7 Parameter uncertainty in model testing

Parameter	Type	Uncertainty
Radiator		
Convective heat transfer area multiplier	Absolute	$Beta(1.3,2)\times 3+3$
TRV curve second-order coefficient	Relative	$Beta(2,1.3)+0.5$
TRV curve first-order coefficient	Relative	$Beta(1.3,2)+0.5$
Radiator rated flow rate multiplier	Absolute	$Beta(2,2)$
Heating loop		
Pipe-to-air convection coefficient (W/m^2K)	Absolute	$Beta(1.3,2)\times 4+2$
Pipe ambient air temperature (°C)	Absolute	$Beta(2,2)\times 8+6$
Pipe insulation thickness (m)	Absolute	$Beta(2,2)\times 0.04+0.02$
Building fabric		
Exterior wall insulation conductivity	Relative	$Beta(2,2)\times 0.8+0.8$
Exterior wall brick density	Relative	$Beta(2,2)\times 0.6+0.8$
Exterior wall brick specific heat	Relative	$Beta(2,2)+0.5$
Exterior glazing equivalent U-value	Relative	$Beta(2,2)+0.5$
Exterior window glazing solar reflectance	Absolute	$Beta(2,2)\times 0.04+0.06$
Exterior window glazing solar transmittance	Absolute	$Beta(2,2)\times 0.3+0.5$
ELA per exterior envelope area (cm^2/m^2)	Absolute	$\text{lognormal}(1.28, 0.88^2)$
Inter-room constant infiltration rate (m^3/s)	Absolute	$Beta(2,2)\times 0.003$
Room thermal mass		
Room air capacity multiplier	Absolute	$Beta(2,2)\times 2.6+1$
Radiant area as a percentage of floor area	Absolute	$Beta(2,2)\times 0.1+0.05$
Room internal load		
Occupant peak load density	Absolute	$Beta(1.3,2)\times 0.8+0.2$
Occupant base load density	Absolute	$Beta(1.3,2)\times 0.8+0.2$
Occupant peak load hours	Absolute	$Beta(2,2)\times 6+7$
Radiant ratio	Absolute	$Beta(2,2)\times 0.3+0.15$
Occupant behavior		
Shading incident radiation threshold (W/m^2)	Absolute	$Beta(1.3,2)\times 500$
Percentage of openable window area	Absolute	$Beta(2,2)\times 0.02$
Radiator TRV setpoint except stair/corridor	Absolute	$Beta(2,2)\times 3+19.5$
Stair/corridor radiator TRV setpoint	Absolute	$Beta(2,2)\times 3+14.5$

Daily heating consumption



Daily meeting room mean air temperature



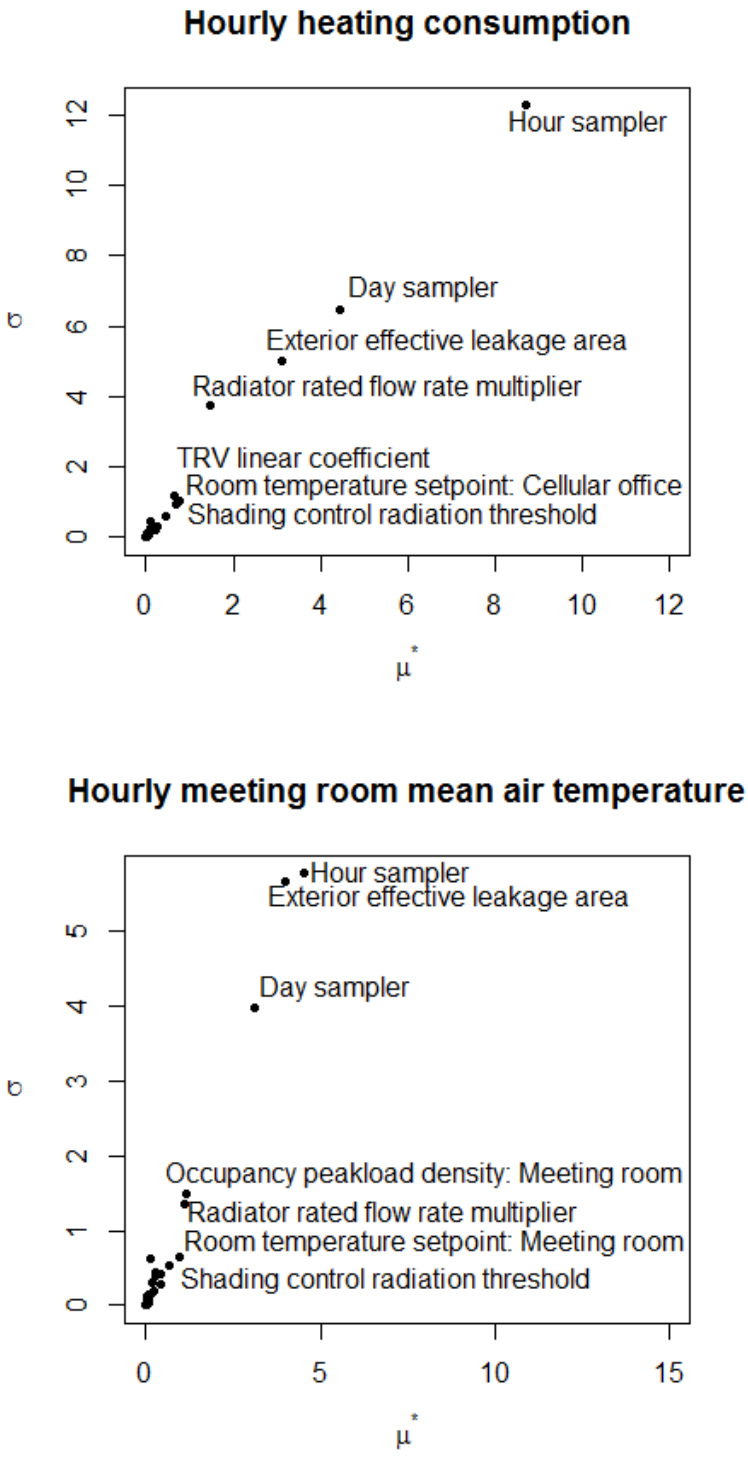


Figure A.7 Full result of parameter screening

Table A.8 Daily heating parameter screening result of the top 20 parameters

Parameter	Main effects	Interactions
Exterior effective leakage area	8.69e-01	1.16e+00
Radiator rated flow rate multiplier	8.32e-01	1.60e+00
Day sampler	6.92e-01	4.11e-01
TRV linear coefficient	5.20e-01	1.69e+00
Shading control radiation threshold	2.34e-01	1.62e-01
Room temperature setpoint: Cellular office	2.33e-01	2.06e-01
Exterior glazing pseudo U-value	1.59e-01	1.26e-01
Room temperature setpoint: Meeting room	6.82e-02	6.24e-02
Room temperature setpoint: Corridor	5.98e-02	3.35e-02
Occupancy peak load density: Meeting room	5.46e-02	5.11e-02
Exterior glazing transmissivity	5.40e-02	4.29e-02
Radiator area multiplier	3.77e-02	1.50e-01
Occupancy peak load density: Cellular office	3.71e-02	4.53e-02
TRV quadratic coefficient	3.17e-02	5.18e-02
Room temperature setpoint: Open office	3.10e-02	2.10e-02
Hot water pipe insulation thickness	2.64e-02	1.82e-02
Exterior insulation conductivity	2.16e-02	1.94e-02
Plug load weekend sampler: Cellular office	2.12e-02	5.61e-02
Lighting weekday sampler: Cellular office	2.05e-02	3.38e-02
Exterior brick heat capacity	2.01e-02	1.55e-02

Table A.9 Daily temperature parameter screening result of the top 20 parameters

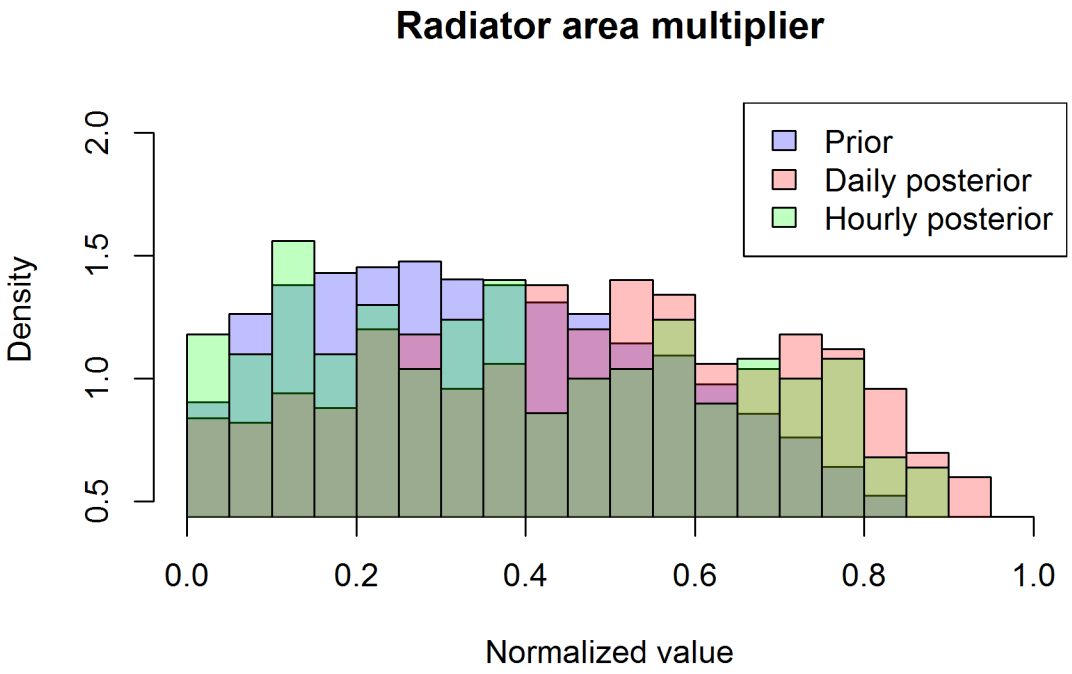
Parameter	Main effects	Interactions
Exterior effective leakage area	4.37e+00	6.19e+00
Day sampler	2.02e+00	2.75e+00
Radiator rated flow rate multiplier	1.19e+00	1.24e+00
Occupancy peak load density: Meeting room	9.81e-01	8.87e-01
Room temperature setpoint: Meeting room	9.16e-01	5.98e-01
Shading control radiation threshold	7.93e-01	5.92e-01
TRV linear coefficient	5.42e-01	8.81e-01
Exterior glazing pseudo U-value	4.70e-01	2.61e-01
Occupancy peak load hours: Meeting room	4.13e-01	4.07e-01
Lighting weekday sampler: Meeting room	2.54e-01	2.62e-01
Occupancy baseload density: Meeting room	2.50e-01	1.52e-01
Exterior glazing transmissivity	1.54e-01	1.50e-01
Plug load weekday sampler: Meeting room	1.46e-01	2.17e-01
TRV quadratic coefficient	7.78e-02	9.17e-02
Exterior brick heat capacity	6.74e-02	2.72e-02
Occupant radiant heat percentage	6.68e-02	6.70e-02
Exterior insulation conductivity	6.21e-02	3.39e-02
Radiator area multiplier	5.62e-02	1.16e-01
Plug load weekend sampler: Meeting room	4.91e-02	1.11e-01
Lighting weekend sampler: Meeting room	3.99e-02	1.01e-01

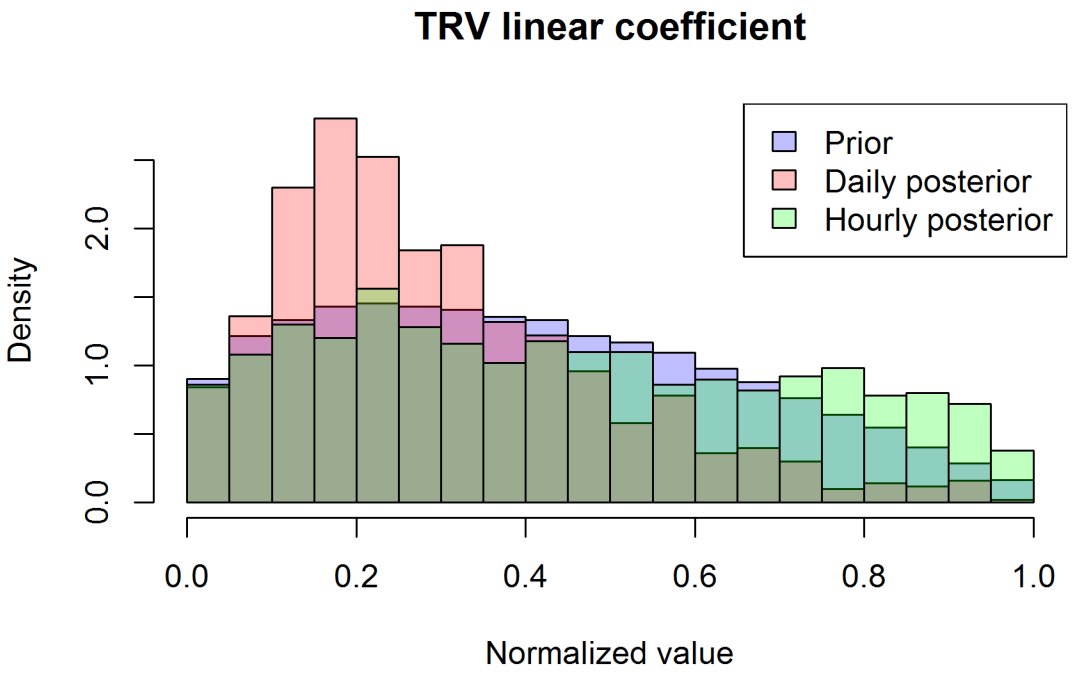
Table A.10 Hourly heating parameter screening result of the top 20 parameters

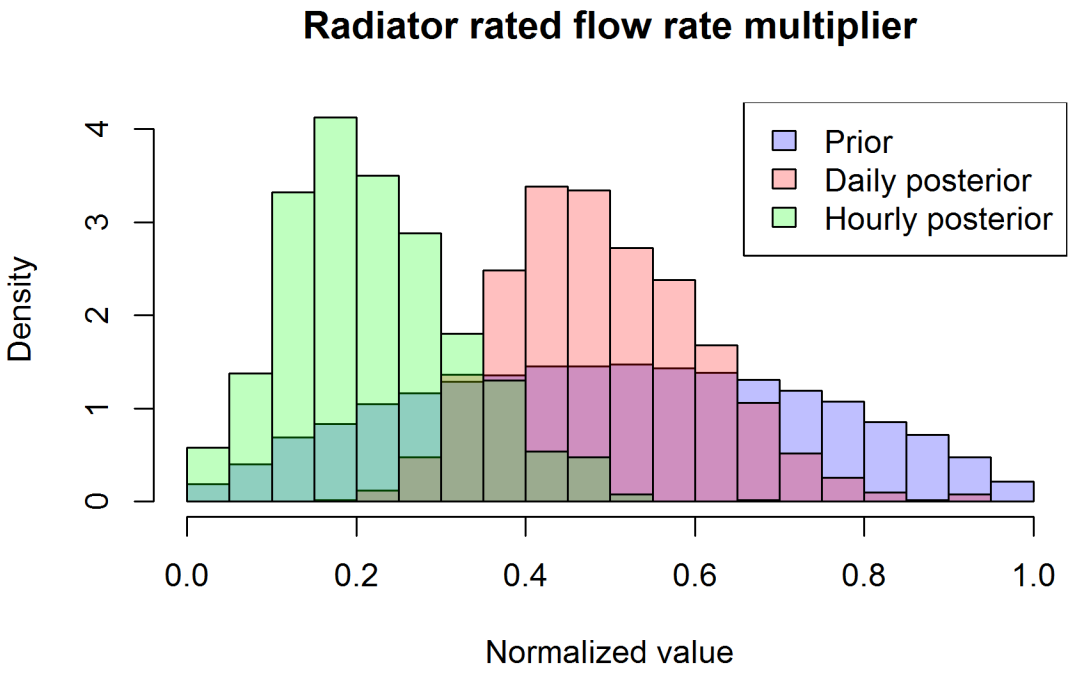
Parameter	Main effects	Interactions
Hour sampler	4.31e+00	5.83e+00
Day sampler	2.30e+00	3.74e+00
Exterior effective leakage area	1.05e+00	1.71e+00
Radiator rated flow rate multiplier	5.00e-01	1.27e+00
Room temperature setpoint: Cellular office	2.63e-01	3.49e-01
Shading control radiation threshold	2.27e-01	3.21e-01
TRV linear coefficient	2.18e-01	4.04e-01
Exterior glazing pseudo U-value	1.55e-01	2.05e-01
Room temperature setpoint: Corridor	8.24e-02	1.01e-01
Room temperature setpoint: Meeting room	6.78e-02	7.94e-02
Occupancy peak load density: Meeting room	6.07e-02	9.58e-02
Exterior glazing transmissivity	5.61e-02	6.93e-02
Occupancy peak load density: Cellular office	4.31e-02	8.80e-02
Plug load weekday sampler: Cellular office	3.58e-02	8.32e-02
Radiator area multiplier	3.42e-02	1.54e-01
Room temperature setpoint: Open office	2.92e-02	5.71e-02
Hot water pipe insulation thickness	2.69e-02	3.31e-02
Plug load weekday sampler: Meeting room	2.60e-02	4.79e-02
Exterior insulation conductivity	2.33e-02	2.79e-02
Lighting weekday sampler: Meeting room	2.22e-02	4.69e-02

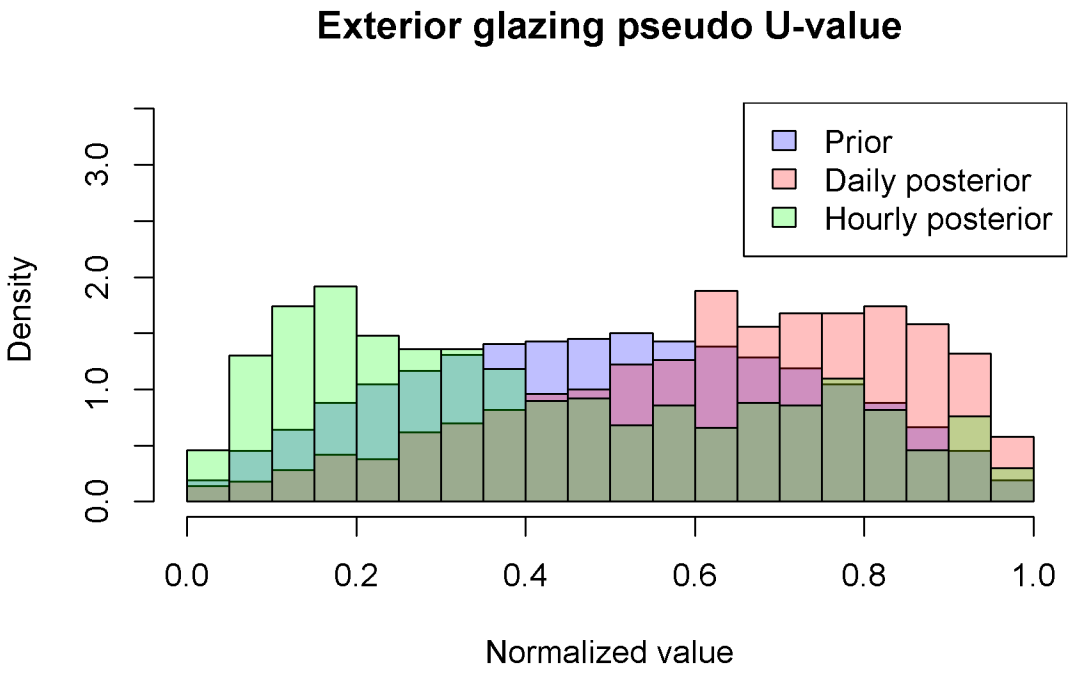
Table A.11 Hourly temperature parameter screening result of the top 20 parameters

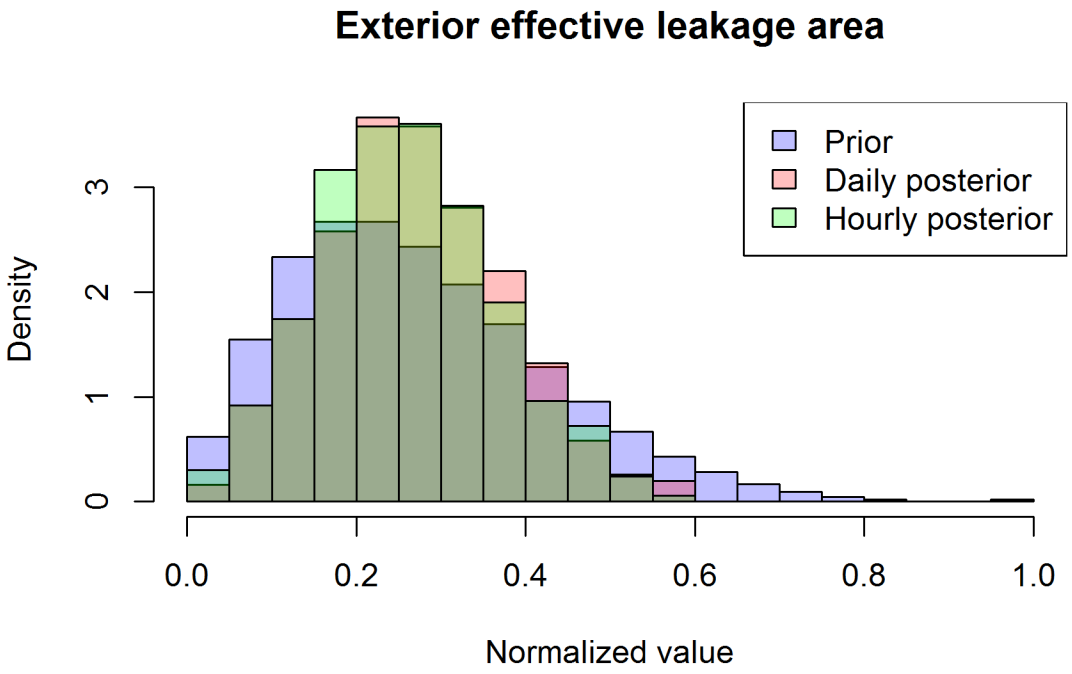
Parameter	Main effects	Interactions
Hour sampler	4.55e+00	5.76e+00
Exterior effective leakage area	3.98e+00	5.65e+00
Day sampler	3.13e+00	3.97e+00
Occupancy peak load density: Meeting room	1.14e+00	1.50e+00
Radiator rated flow rate multiplier	1.10e+00	1.36e+00
Room temperature setpoint: Meeting room	9.43e-01	6.42e-01
Shading control radiation threshold	6.38e-01	5.36e-01
TRV linear coefficient	4.37e-01	4.21e-01
Exterior glazing pseudo U-value	4.35e-01	2.79e-01
Lighting weekday sampler: Meeting room	2.89e-01	4.02e-01
Occupancy peak load hours: Meeting room	2.71e-01	4.49e-01
Occupancy baseload density: Meeting room	2.32e-01	1.97e-01
Plug load weekday sampler: Meeting room	1.88e-01	3.02e-01
Exterior glazing transmissivity	1.31e-01	1.44e-01
Window opening percentage: Meeting room	1.13e-01	6.20e-01
Exterior brick heat capacity	6.80e-02	3.42e-02
Radiator area multiplier	6.08e-02	1.37e-01
TRV quadratic coefficient	6.06e-02	6.66e-02
Lighting weekend sampler: Meeting room	5.64e-02	1.23e-01
Occupant radiant heat percentage	5.54e-02	9.59e-02

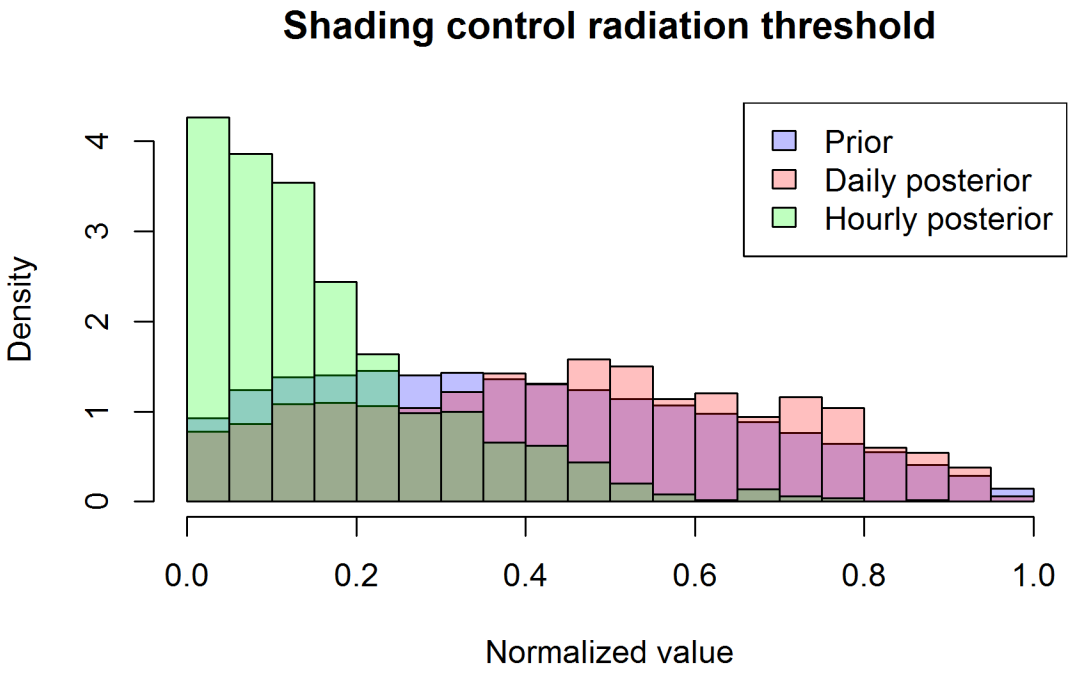

Radiator area multiplier
Prior
Daily posterior
Hourly posterior
Density
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value

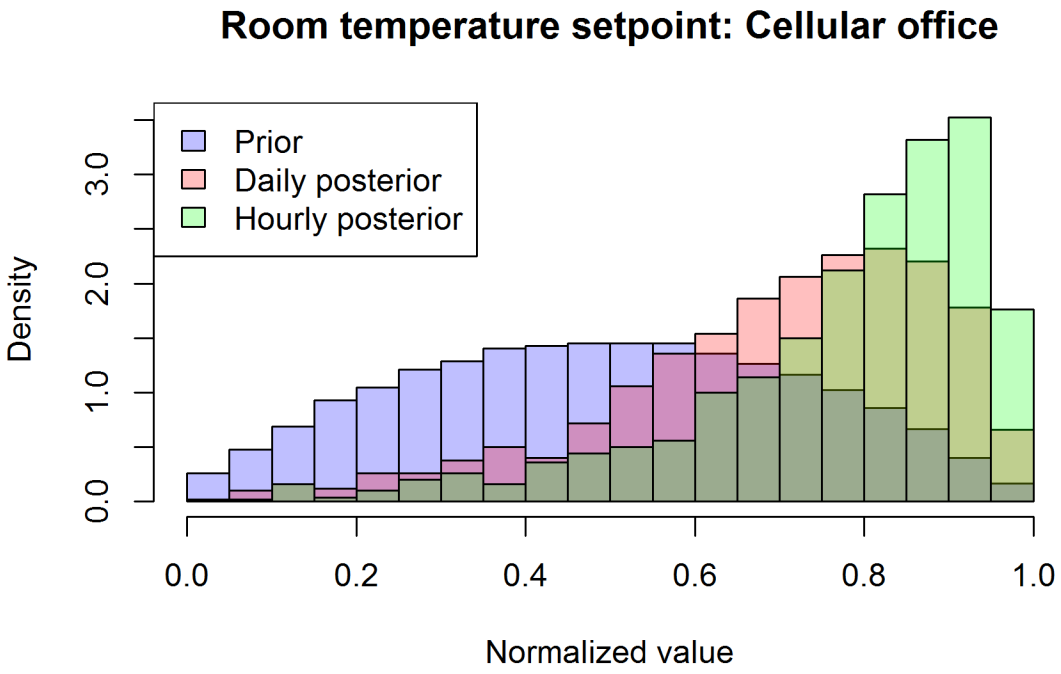

TRV linear coefficient
Prior
Daily posterior
Hourly posterior
Density
0.0
1.0
2.0
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value

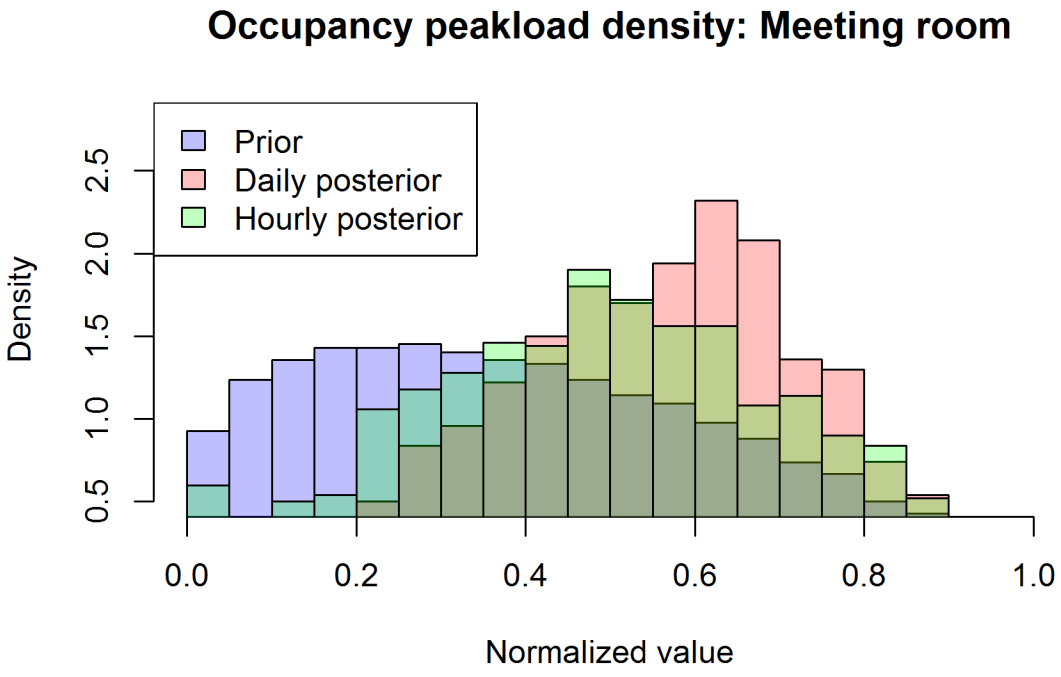

Radiator rated flow rate multiplier
Prior
Daily posterior
Hourly posterior
Density
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value

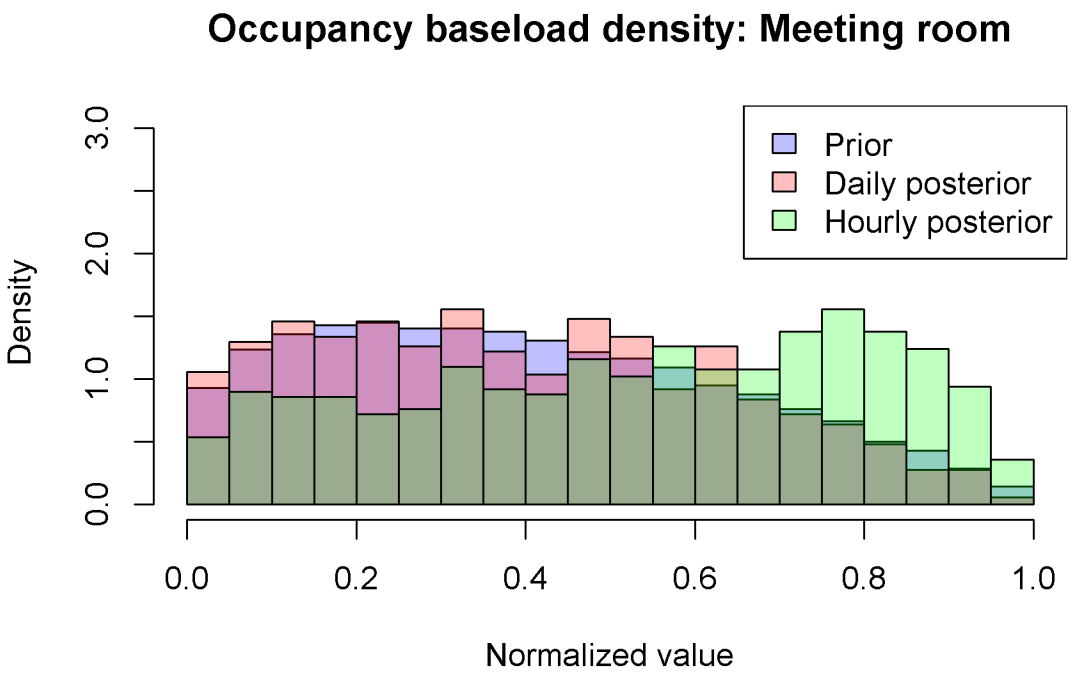

Exterior glazing pseudo U-value
Prior
Daily posterior
Hourly posterior
Density
0.0
1.0
2.0
3.0
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value

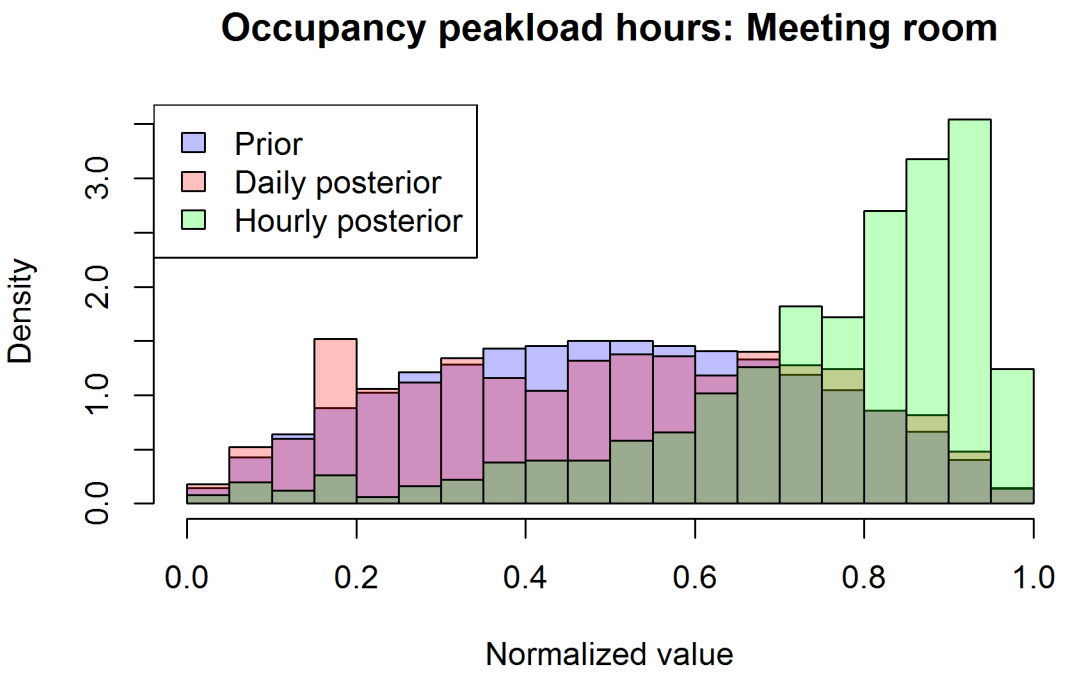

Exterior effective leakage area
Prior
Daily posterior
Hourly posterior
Density
0
1
2
3
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value


Shading control radiation threshold
Prior
Daily posterior
Hourly posterior
Density
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value


Room temperature setpoint: Cellular office
Prior
Daily posterior
Hourly posterior
Density
0.0
1.0
2.0
3.0
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value


Occupancy peakload density: Meeting room
Prior
Daily posterior
Hourly posterior
Density
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value


Occupancy baseload density: Meeting room
Prior
Daily posterior
Hourly posterior
Density
0.0
1.0
2.0
3.0
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value


Occupancy peakload hours: Meeting room
Prior
Daily posterior
Hourly posterior
Density
0.0
1.0
2.0
3.0
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value

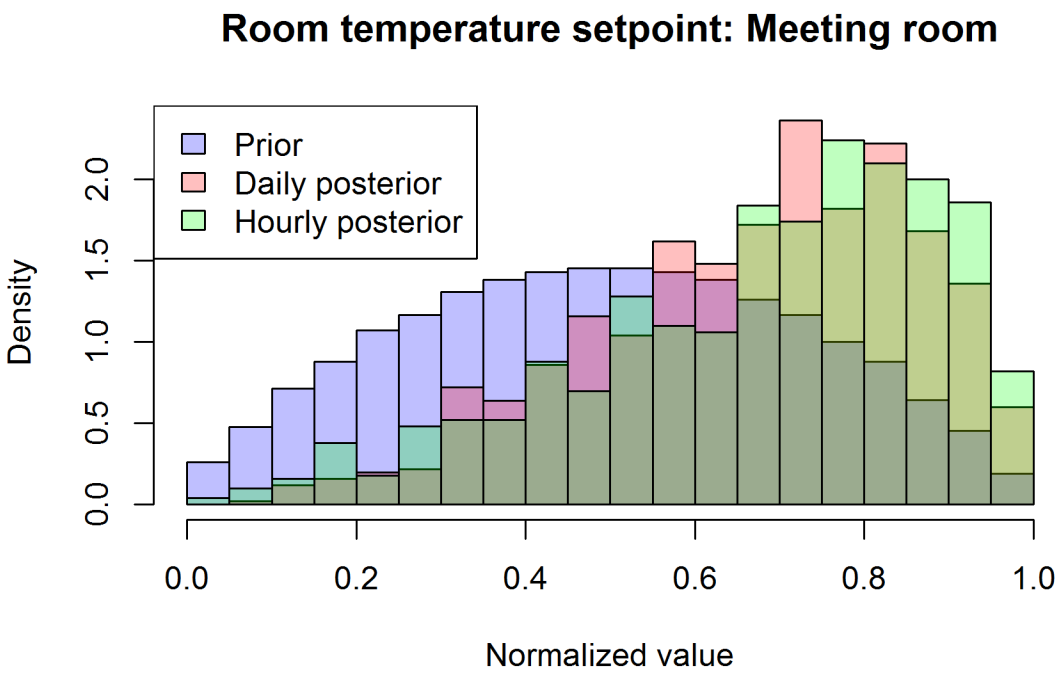
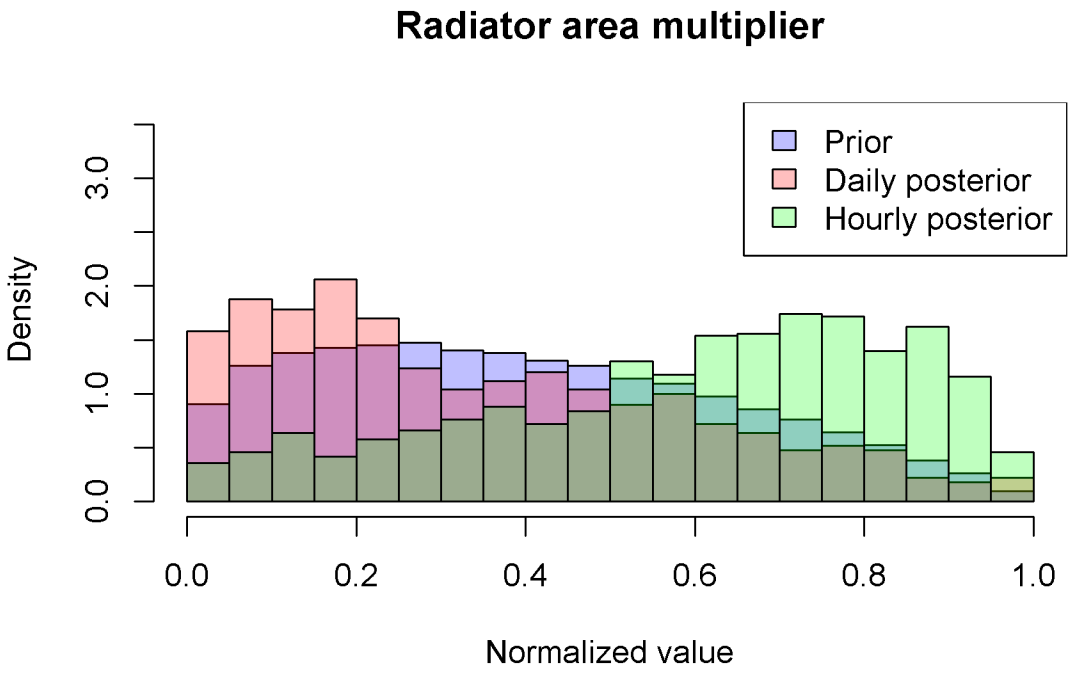
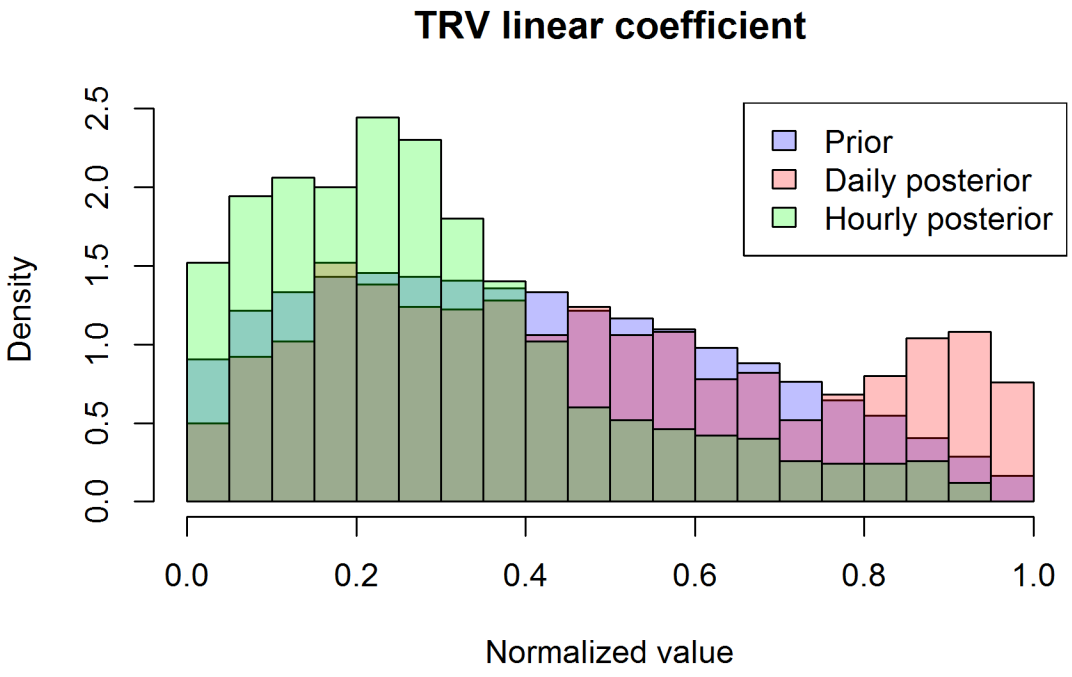
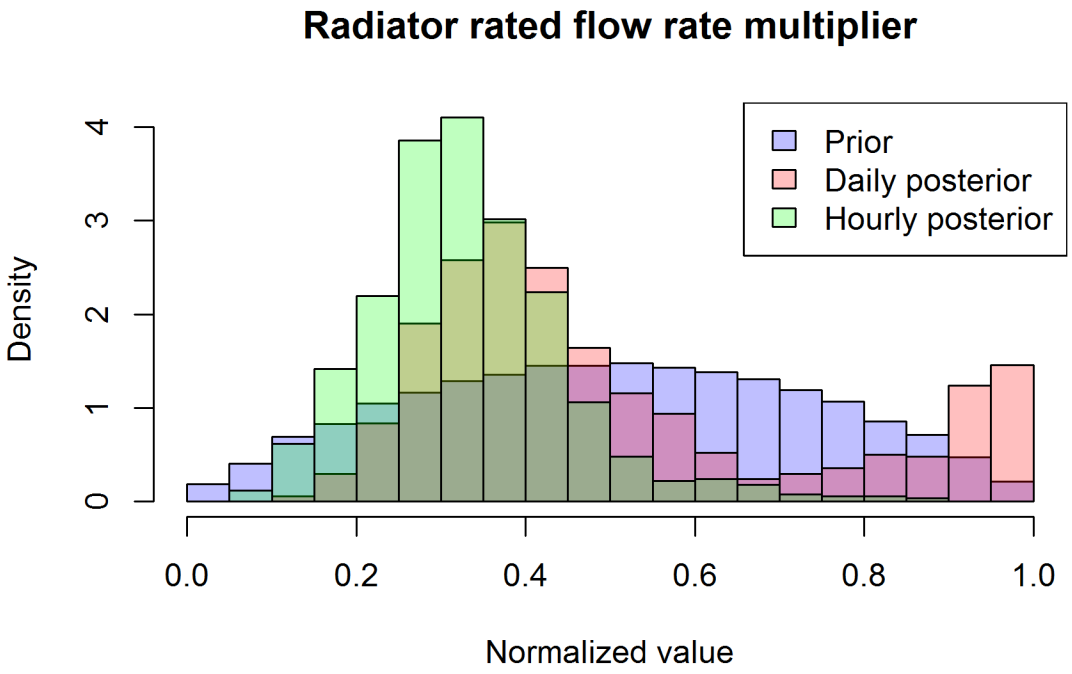
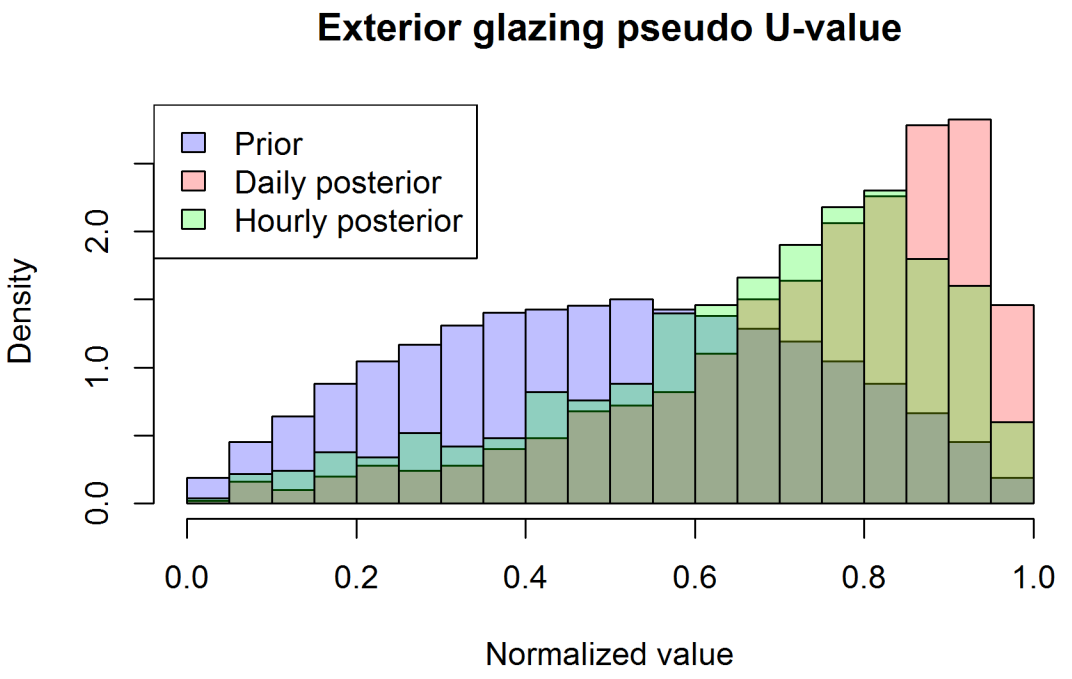


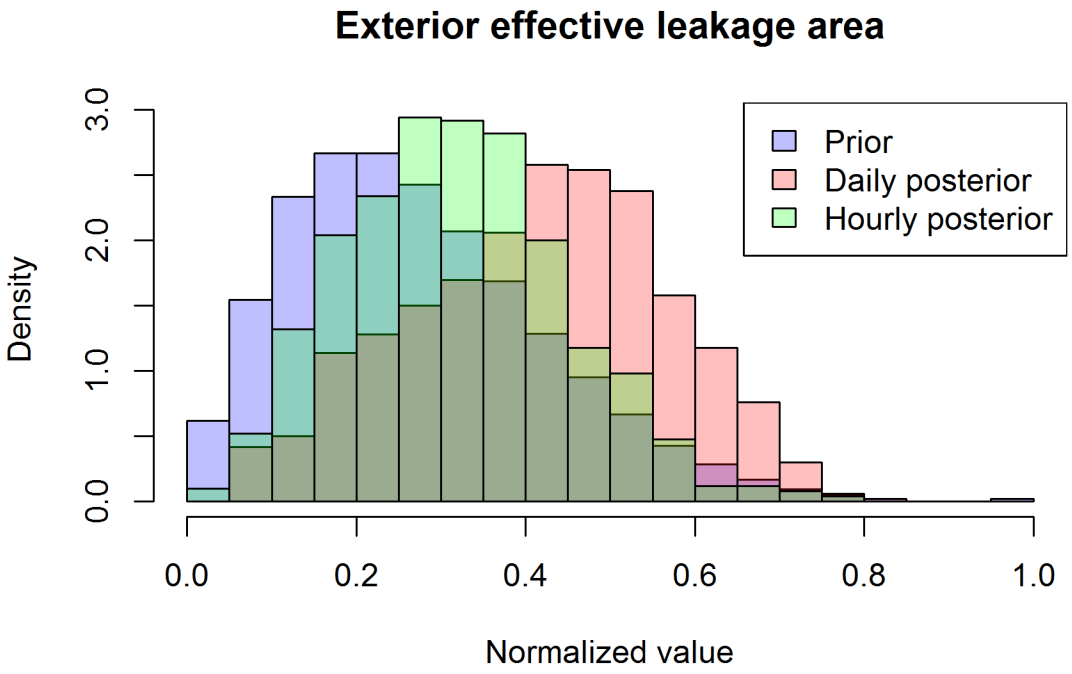
Figure A.8 D/H-BI prior and posterior estimates of calibration parameters

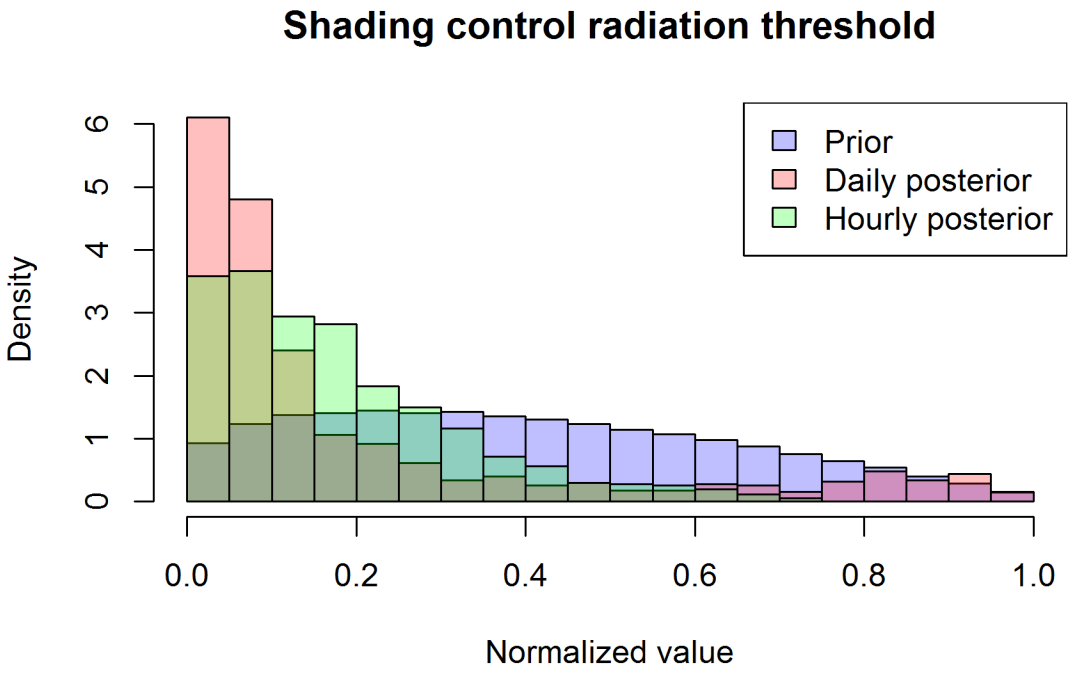

Radiator area multiplier
Prior
Daily posterior
Hourly posterior
Density
0.0
1.0
2.0
3.0
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value


TRV linear coefficient
Prior
Daily posterior
Hourly posterior
Density
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value


Radiator rated flow rate multiplier
Prior
Daily posterior
Hourly posterior
Density
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value


Exterior glazing pseudo U-value
Prior
Daily posterior
Hourly posterior
Density
0.0
1.0
2.0
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value


Exterior effective leakage area
Prior
Daily posterior
Hourly posterior
Density
0.0
1.0
2.0
3.0
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value


Shading control radiation threshold
Prior
Daily posterior
Hourly posterior
Density
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value

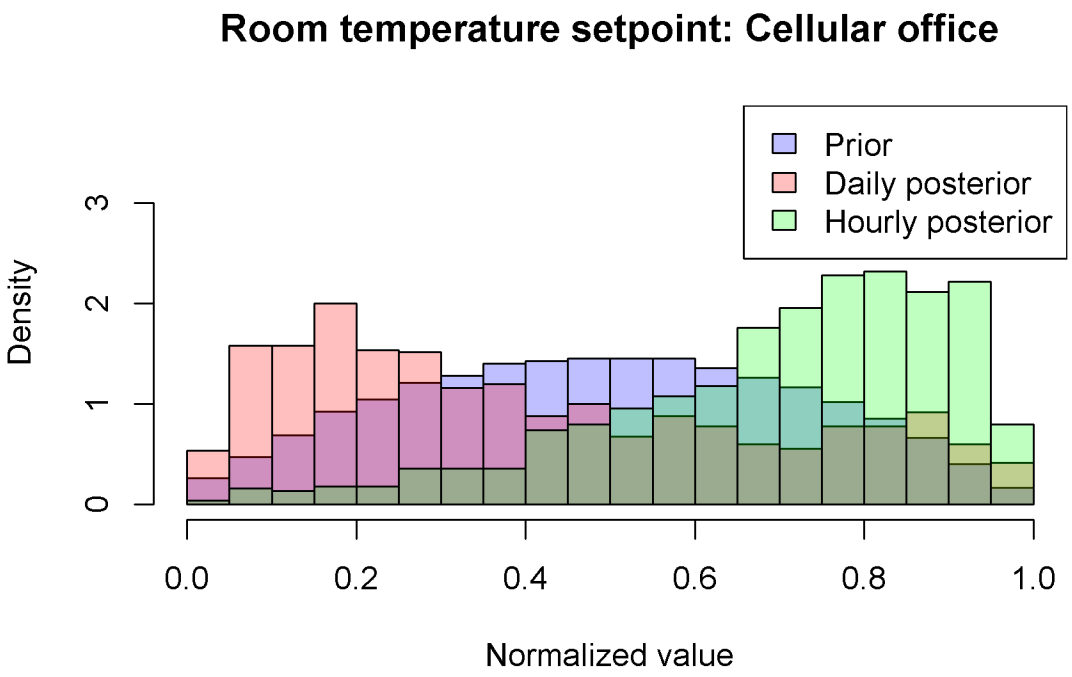
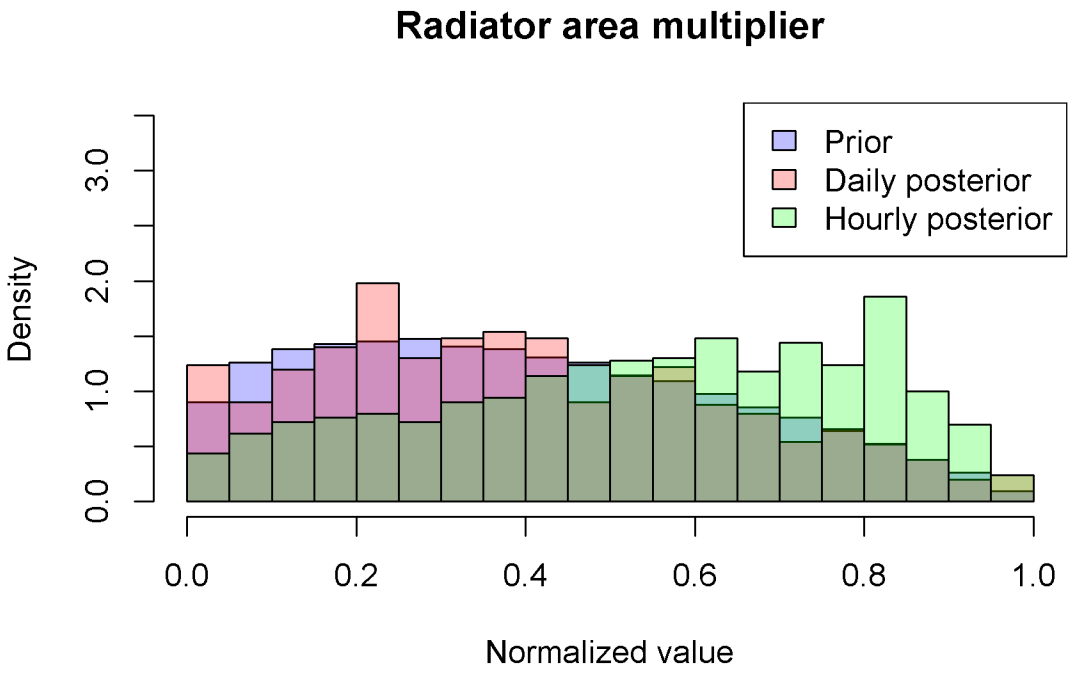
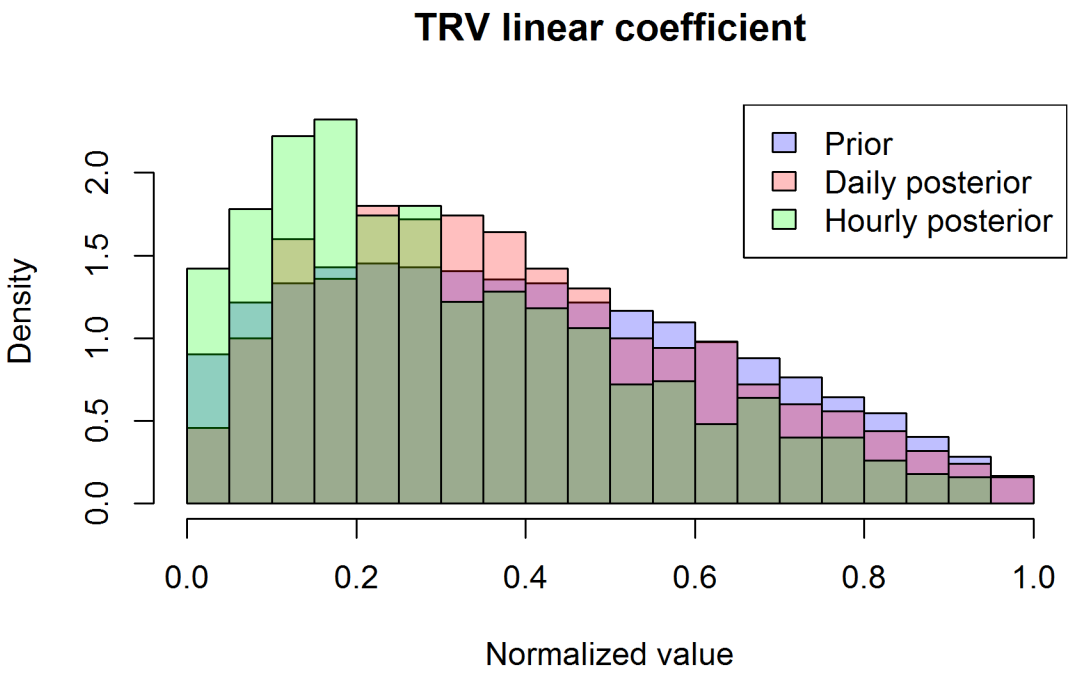
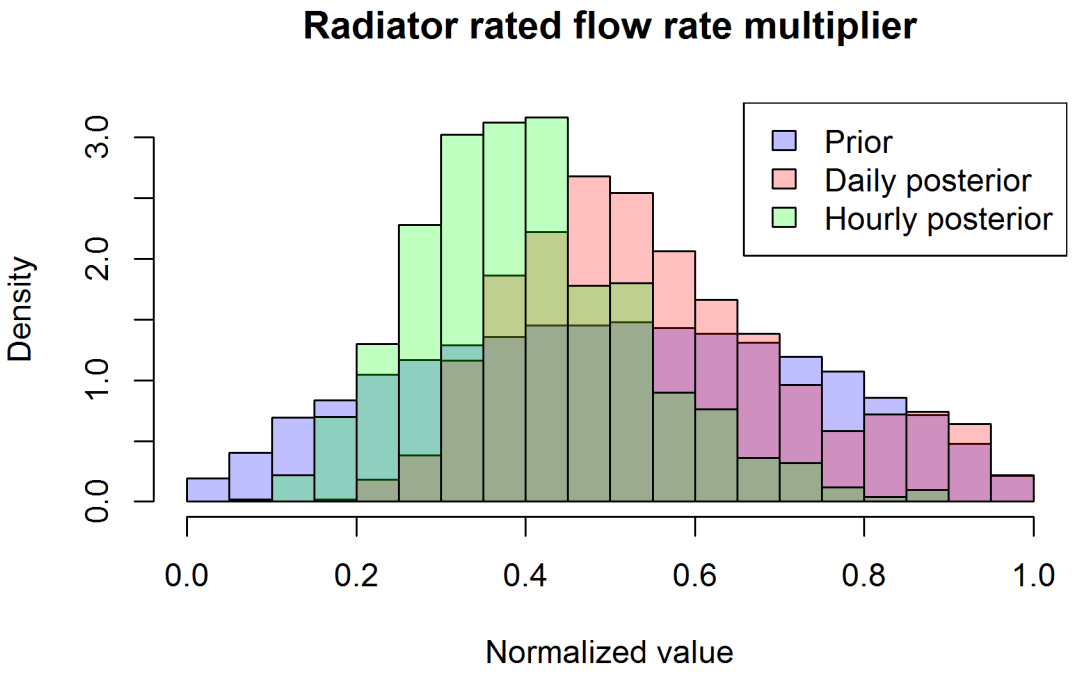
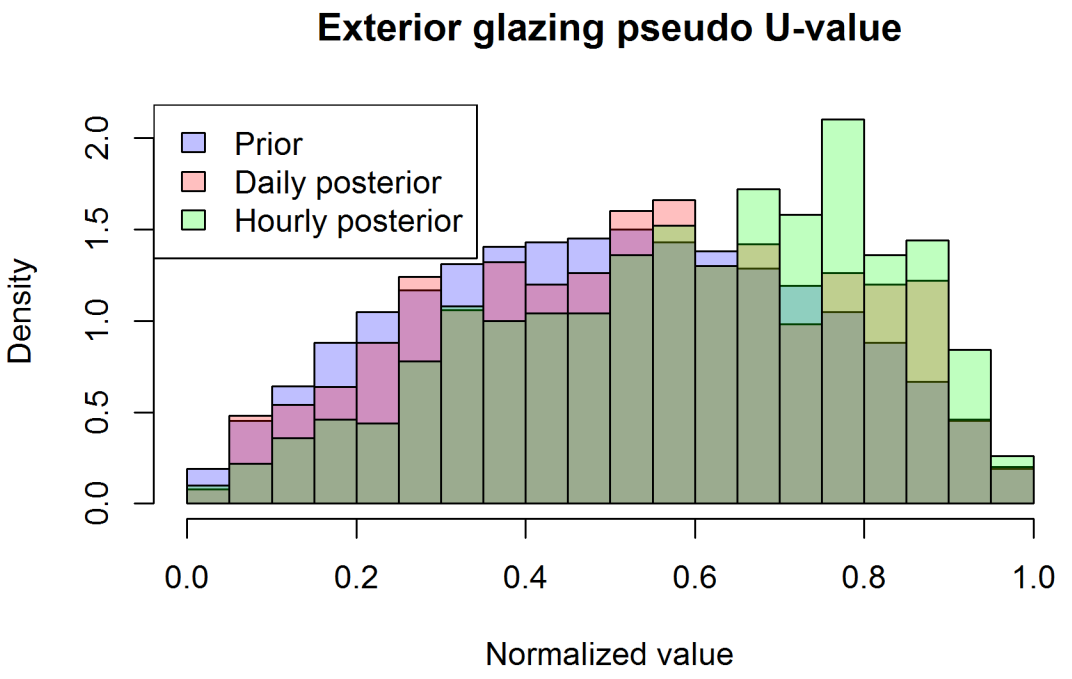


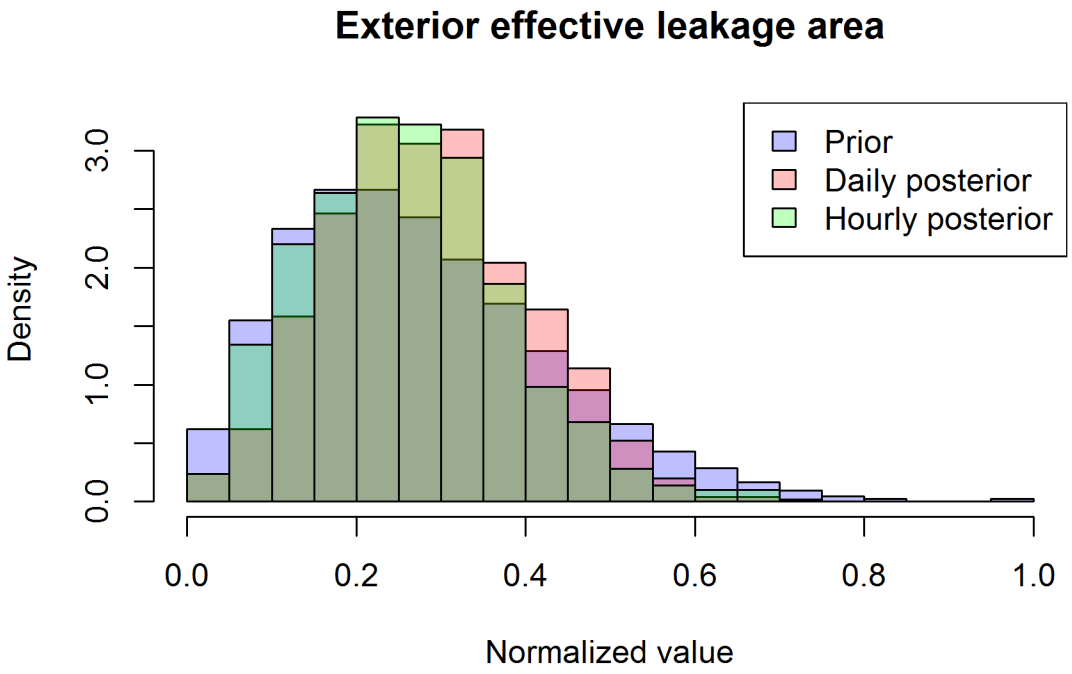
Figure A.9 D/H-UI prior and posterior estimates of calibration parameters

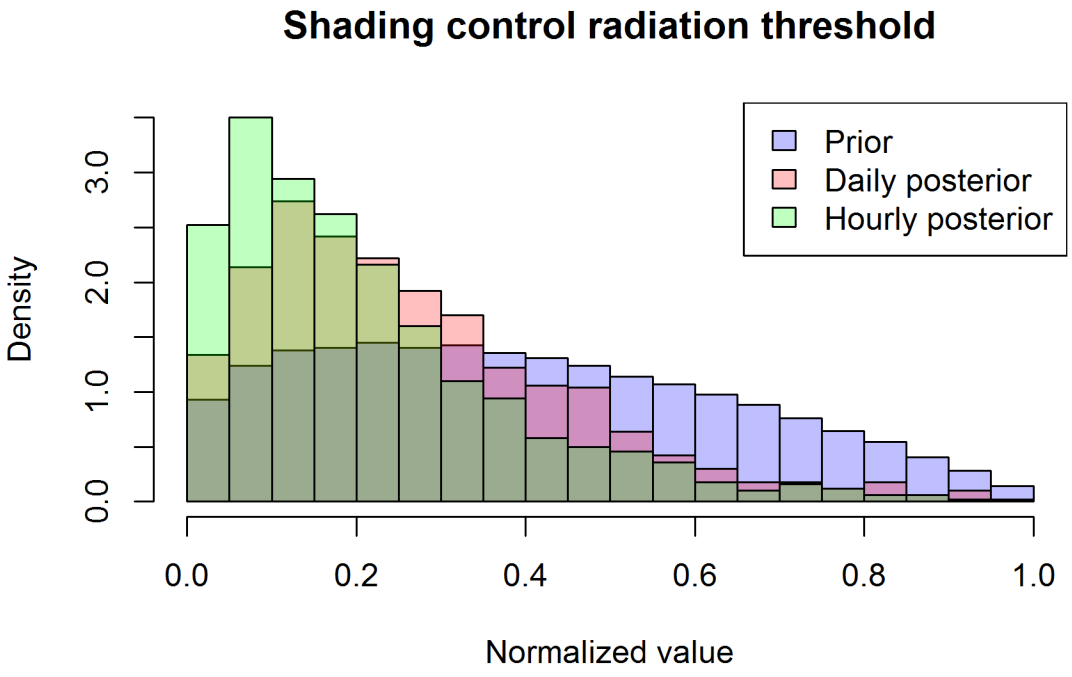

Radiator area multiplier
Prior
Daily posterior
Hourly posterior
Density
0.0
1.0
2.0
3.0
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value


TRV linear coefficient
Prior
Daily posterior
Hourly posterior
Density
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value


Radiator rated flow rate multiplier
Prior
Daily posterior
Hourly posterior
Density
0.0
1.0
2.0
3.0
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value


Exterior glazing pseudo U-value
Prior
Daily posterior
Hourly posterior
Density
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value


Exterior effective leakage area
Prior
Daily posterior
Hourly posterior
Density
0.0
1.0
2.0
3.0
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value


Shading control radiation threshold
Prior
Daily posterior
Hourly posterior
Density
0.0
1.0
2.0
3.0
0.0
0.2
0.4
0.6
0.8
1.0
Normalized value

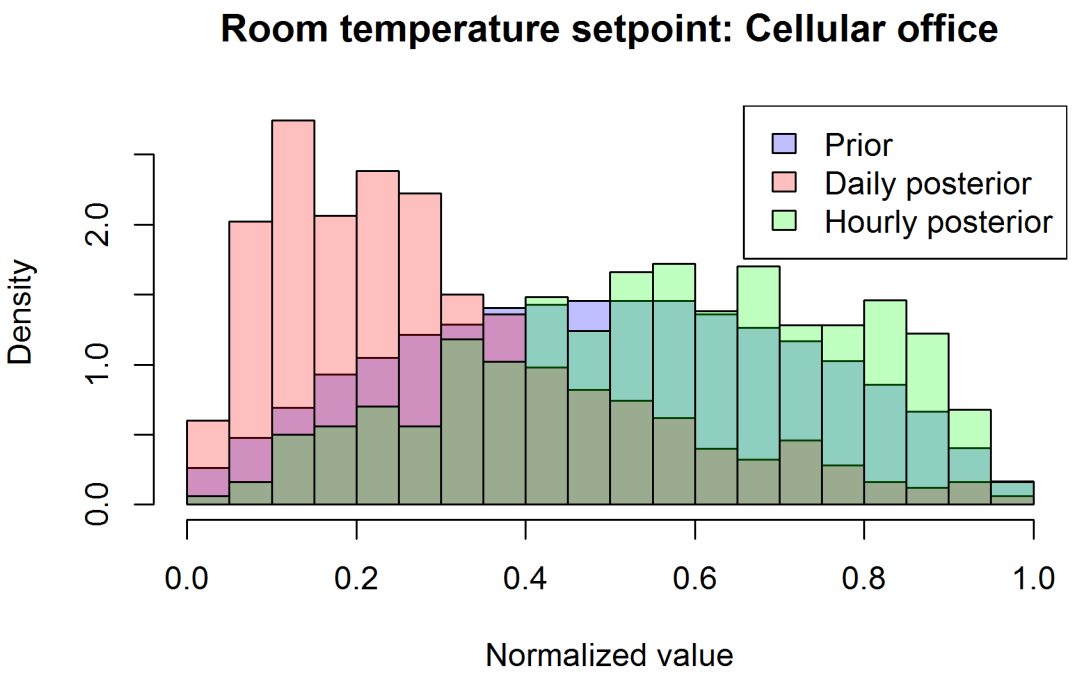


Figure A.10 D/H-EI prior and posterior estimates of calibration parameters

REFERENCES

ASHRAE. (2002). ASHRAE Guideline 14-2002: Measurement of energy and demand savings. *American Society of Heating, Ventilating, and Air Conditioning*.

Bal, G., Langmore, I., & Marzouk, Y. (2013). Bayesian inverse problems with Monte Carlo forward models. *Inverse Problems and Imaging*, *7*(1), 81–105. https://doi.org/10.3934/ipi.2013.7.81

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. a, Cavendish, J., … Tu, J. (2007). A framework for validation of computer models. *Technometrics*, *49*(2), 138–154. https://doi.org/10.1198/004017007000000092

Booth, A. T., Choudhary, R., & Spiegelhalter, D. J. (2012). Handling uncertainty in housing stock models. *Building and Environment*, *48*(1), 35–47. https://doi.org/10.1016/j.buildenv.2011.08.016

Booth, A. T., Choudhary, R., & Spiegelhalter, D. J. (2013). A hierarchical Bayesian framework for calibrating micro-level models with macro-level data. *Journal of Building Performance Simulation*, *6*(4), 293–318. https://doi.org/10.1080/19401493.2012.723750

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791. https://doi.org/10.2307/2286841

Bronson, D. J., Hinchey, S. B., Haberl, J. S., & O'Neal, D. L. (1992). A procedure for calibrating the DOE-2 simulation program to non-weather-dependent measured loads. In M. Geshwiler (Ed.), *ASHRAE Transactions* (Vol. 98, pp. 636–652). Atlanta: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

Brynjarsdottir, J., & O'Hagan, A. (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse Problems*, *xx*, 1–21. https://doi.org/10.1088/0266-5611/30/11/114007

Campolongo, F., Cariboni, J., & Saltelli, A. (2007). An effective screening design for sensitivity analysis of large models. *Environmental Modelling & Software*, *22*(10), 1509–1518. https://doi.org/http://doi.org/10.1016/j.envsoft.2006.10.004

Carroll, W. L., & Hitchcock, R. J. (1993). Tuning simulated building descriptions to match actual utility data - methods and implementation. *ASHRAE Transactions*, *99*(2), 928–934.

Chaudhary, G., New, J., Sanyal, J., Im, P., O'Neill, Z., & Garg, V. (2016). Evaluation of "Autotune" calibration against manual calibration of building energy models. *Applied Energy*, *182*, 115–134. https://doi.org/10.1016/j.apenergy.2016.08.073

Chong, A., & Lam, K. P. (2015). Uncertainty analysis and parameter estimation of HVAC systems in building energy models. In *Proceedings of the 14th International Conference of the International Building Performance Simulation Association* (pp. 2788–2795).

Clarke, J., Strachan, P. A., & Pernot, C. (1993). An approach to the calibration of building energy simulation models. *Transitions-American Society of Heating Refrigerating and Air Conditioning Engineers*, 917–930.

Coakley, D., Raftery, P., & Keane, M. (2014). A review of methods to match building energy simulation models to measured data. *Renewable and Sustainable Energy Reviews*, *37*, 123–141. https://doi.org/10.1016/j.rser.2014.05.007

Coakley, D., Raftery, P., Molloy, P., & White, G. (2011). Calibration of a detailed BES model to measured data using an evidence-based analytical optimisation approach. *Proceedings of the 12th International Conference of the International Building Performance Simulation Association*, 374–381.

Conti, S., & O'Hagan, A. (2010). Bayesian emulation of complex multi-output and dynamic computer models. *Journal of Statistical Planning and Inference*, *140*(3), 640–651. https://doi.org/10.1016/j.jspi.2009.08.006

Crawley, D. B., Pedersen, C. O., Lawrie, L. K., & Winkelmann, F. C. (2000). EnergyPlus: Energy simulation program. *ASHRAE Journal*, *42*(4).

Deru, M., Field, K., Studer, D., Benne, K., Griffith, B., Torcellini, P., … Crawley, D. (2011). *U.S. Department of Energy commercial reference building models of the national building stock*.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Computing Surveys*, *27*(3), 326–327. https://doi.org/10.1145/212094.212114

Diggle, P. J., & Gratton, R. J. (1984). Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *46*(2), 193–227.

Djuric, N., Novakovic, V., & Frydenlund, F. (2008). Heating system performance estimation using optimization tool and BEMS data. *Energy and Buildings*, *40*(8), 1367–1376. https://doi.org/10.1016/j.enbuild.2007.12.006

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, *195*(2), 216–222. https://doi.org/10.1016/0370-2693(87)91197-X

EIA. (2017). *Annual energy outlook 2017 with projections to 2050*.

Fabrizio, E., & Monetti, V. (2015). Methodologies and advancements in the calibration of building energy models. *Energies*, *8*(4), 2548–2574. https://doi.org/10.3390/en8042548

Fumo, N. (2014). A review on the basics of building energy estimation. *Renewable and Sustainable Energy Reviews*, *31*, 53–60. https://doi.org/10.1016/j.rser.2013.11.040

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (Third Edit). CRC Press.

Gestwick, M. J., & Love, J. a. (2014). Trial application of ASHRAE 1051-RP: calibration method for building energy simulation. *Journal of Building Performance Simulation*, *7*(5), 346–359. https://doi.org/10.1080/19401493.2013.838698

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *69*(2), 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

Gneiting, T., & Katzfuss, M. (2014). Probabilistic Forecasting. *Annual Review of Statistics and Its Application*, *1*, 125–151. https://doi.org/10.1146/annurev-statistics-062713-085831

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Guillas, S., Rougier, J., Maute, A., Richmond, A. D., & Linkletter, C. D. (2009). Bayesian calibration of the thermosphere-Ionosphere electrodynamics general circulation model (TIE-GCM). *Geoscientific Model Development Discussions*, *2*(1), 485–506. https://doi.org/10.5194/gmdd-2-485-2009

Haberl, J. S., & Bou-Saada, T. E. (1998). Procedures for calibrating hourly simulation models to measured building energy and environmental data. *Journal of Solar Energy Engineering*, *120*(May 1998), 193. https://doi.org/10.1115/1.2888069

Haberl, J., Sparks, R., & Culp, C. (1996). Exploring new techniques for displaying complex building energy consumption data. *Energy and Buildings*, *24*(1), 27–38. https://doi.org/10.1016/0378-7788(95)00959-0

Hamill, T. (2007). Common verification methods for ensemble forecasts. Retrieved from http://slideplayer.com/slide/9514952/

Heo, Y. (2011). *Bayesian calibration of building energy models for energy retrofit decision-making under uncertainty*. Georgia Institute of Technology.

Heo, Y., Choudhary, R., & Augenbroe, G. a. (2012). Calibration of building energy models for retrofit analysis under uncertainty. *Energy and Buildings*, *47*, 550–560. https://doi.org/10.1016/j.enbuild.2011.12.029

Heo, Y., Graziano, D. J., Guzowski, L., & Muehleisen, R. T. (2015). Evaluation of calibration efficacy under different levels of uncertainty. *Journal of Building Performance Simulation*, *8*(3), 135–144. https://doi.org/10.1080/19401493.2014.896947

Heo, Y., & Zavala, V. M. (2012). Gaussian process modeling for measurement and verification of building energy savings. *Energy and Buildings*, *53*, 7–18. https://doi.org/10.1016/j.enbuild.2012.06.024

Hersbach, H. (2000). Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, *15*(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo, J. a., & Ryne, R. D. (2004). Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, *26*(2), 448–466. https://doi.org/10.1137/S1064827503426693

Hopfe, C. J. (2009). *Uncertainty and sensitivity analysis in building performance simulation for decision support and design optimization*. Eindhoven University.

IPMVP. (2002). Concepts and options for determining energy and water savings. *International Performance Measurement & Verification Protocol*, *1*(DOE/GO-102002-1554), 1–93. https://doi.org/DOE/GO-102002-1554

Jensen, S. O. (1993). *Validation of building energy simulation programs, Part I and II.* Brussels.

Judkoff, R., & Neymark, J. (2006). Model validation and testing: the methodological foundation of ASHRAE Standard 140. In *the ASHRAE 2006 Annual Meeting*. Quebec City, Canada.

Judkoff, R., Polly, B., Bianchi, M., & Neymark, J. (2010). *Building Energy Simulation Test for Existing Homes (BESTEST-EX).*

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(3), 425–464. https://doi.org/10.1111/1467-9868.00294

Lavigne, K. (2009). Assisted calibration in building simulation–algorithm description and case studies. In *Proceedings of the 11th International Conference of the International Building Performance Simulation Association* (pp. 1498–1505).

Li, Q., Augenbroe, G., & Brown, J. (2016). Assessment of linear emulators in lightweight Bayesian calibration of dynamic building energy models for parameter estimation and performance prediction. *Energy and Buildings*, *124*, 194–202. https://doi.org/10.1016/j.enbuild.2016.04.025

Li, Q., Gu, L., Augenbroe, G., Wu, C. F. J., & Brown, J. (2015). A generic approach to calibrate building energy models under uncertainty using Bayesian inference. In *Proceedings of the 14th International Conference of the International Building Performance Simulation Association* (pp. 2947–2955). Hyderabad, India.

Liu, G., & Liu, M. (2011). A rapid calibration procedure and case study for simplified simulation models of commonly used HVAC systems. *Building and Environment*, *46*(2), 409–420. https://doi.org/10.1016/j.buildenv.2010.08.002

Liu, M., & Claridge, D. E. (1998). Use of calibrated HVAC system models to optimize system operation. *Journal of Solar Energy Engineering*, *120*(2), 131. https://doi.org/10.1115/1.2888056

Liu, M., Liu, G., Claridge, D., & Haberl, J. (2006). *Development of procedures to determine in-situ performance of commonly used HVAC systems*.

Liu, M., Song, L., Wei, G., & Claridge, D. E. (2004). Simplified building and air-handling unit model calibration and applications. *Journal of Solar Energy Engineering*, *126*(1), 601–609. https://doi.org/10.1115/1.1639380

Lomas, K. J., Eppel, H., Martin, C. J., & Bloomfield, D. P. (1997). Empirical validation of building energy simulation programs. *Energy and Buildings*, *26*(3), 253–275. https://doi.org/10.1016/S0378-7788(97)00007-8

Ma, Z., Cooper, P., Daly, D., & Ledo, L. (2012). Existing building retrofits: Methodology and state-of-the-art. *Energy and Buildings*, *55*, 889–902. https://doi.org/10.1016/j.enbuild.2012.08.018

Macdonald, I. (2002). *Quantifying the effects of uncertainty in building simulation*. University of Strathclyde.

Machairas, V., Tsangrassoulis, A., & Axarli, K. (2014). Algorithms for optimization of building design: A review. *Renewable and Sustainable Energy Reviews*, *31*(1364), 101–112. https://doi.org/10.1016/j.rser.2013.11.036

Manfren, M., Aste, N., & Moshksar, R. (2013). Calibration and uncertainty analysis for computer models - A meta-model based approach for integrated building energy simulation. *Applied Energy*, *103*, 627–641. https://doi.org/10.1016/j.apenergy.2012.10.031

Menberg, K., Heo, Y., & Choudhary, R. (2016). Sensitivity analysis methods for building energy models: Comparing computational costs and extractable information. *Energy and Buildings*, *133*, 433–445. https://doi.org/10.1016/j.enbuild.2016.10.005

Metropolis, N. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087. https://doi.org/10.1063/1.1699114

Monfet, D., Charneux, R., Zmeureanu, R., & Lemire, N. (2009). Calibration of a building

energy model using measured data. *ASHRAE Transactions*, *115*(1), 348–359.

Morgan, M. G., Dowlatabadi, H., Henrion, M., Keith, D., Lempert, R., McBride, S., … Wilbanks, T. (2009). *Best practice approaches for characterizing, communicating, and incorporating scientific uncertainty in decisionmaking*. Washington, DC, US.

Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, *33*(2), 161–174. https://doi.org/10.2307/1269043

Mustafaraj, G., Marini, D., Costa, A., & Keane, M. (2014). Model calibration for building energy efficiency simulation. *Applied Energy*, *130*, 72–85. https://doi.org/10.1016/j.apenergy.2014.05.019

Nassiopoulos, A., Kuate, R., & Bourquin, F. (2014). Calibration of building thermal models using an optimal control approach. *Energy and Buildings*, *76*, 72–80. https://doi.org/10.1016/j.enbuild.2014.02.052

Nguyen, A.-T., Reiter, S., & Rigo, P. (2014). A review on simulation-based optimization methods applied to building performance analysis. *Applied Energy*, *113*, 1043–1058. https://doi.org/10.1016/j.apenergy.2013.08.061

O'Neill, Z., & Eisenhower, B. (2013). Leveraging the analysis of parametric uncertainty for building energy model calibration. *Building Simulation*, 1–13. https://doi.org/10.1007/s12273-013-0125-8

O'Neill, Z., Eisenhower, B., Yuan, S., Bailey, T., Narayanan, S., & Fonoberov, V. (2011). Modeling and calibration of energy models for a DoD building. *ASHRAE Transactions*, *117*(2), 358–366.

Palomo, E., Marco, J., & Madsem, H. (1991). Method to compare measurements and simulations. In *Proceedings of the 5th International Conference of the International Building Performance Simulation Association*.

Palomo del Barrio, E., & Guyon, G. (2003). Theoretical basis for empirical model validation using parameters space analysis tools. *Energy and Buildings*, *35*(10), 985–996. https://doi.org/10.1016/S0378-7788(03)00038-0

Palomo del Barrio, E., & Guyon, G. (2004). Application of parameters space analysis tools for empirical model validation. *Energy and Buildings*, *36*(1), 23–33.

https://doi.org/10.1016/S0378-7788(03)00039-2

Pan, Y., Huang, Z., & Wu, G. (2007). Calibrated building energy simulation and its application in a high-rise commercial building in Shanghai. *Energy and Buildings*, *39*(6), 651–657. https://doi.org/10.1016/j.enbuild.2006.09.013

Plumlee, M. (2016). Bayesian calibration of inexact computer models. *Journal of the American Statistical Association*, *1459*(April), 1–14. https://doi.org/10.1080/01621459.2016.1211016

Raftery, P., & Keane, M. (2011). Visualising patterns in building performance data. In *Proceedings of the 12th International Conference of the International Building Performance Simulation Association*.

Raftery, P., Keane, M., & Costa, A. (2011). Calibrating whole building energy models: Detailed case study using hourly measured data. *Energy and Buildings*, *43*(12), 3666–3679. https://doi.org/10.1016/j.enbuild.2011.09.039

Raftery, P., Keane, M., & O'Donnell, J. (2011). Calibrating whole building energy models: An evidence-based methodology. *Energy and Buildings*, *43*(9), 2356–2364. https://doi.org/10.1016/j.enbuild.2011.05.020

Ramos Ruiz, G., Fernández Bandera, C., Gómez-Acebo Temes, T., & Sánchez-Ostiz Gutierrez, A. (2016). Genetic algorithm for building envelope calibration. *Applied Energy*, *168*, 691–705. https://doi.org/10.1016/j.apenergy.2016.01.075

Reddy, A., & Maor, I. (2006). *Procedures for reconciling computer-calculated results: ASHRAE Research Project 1051- RP*.

Reddy, T. A. (2006). Literature review on calibration of building energy simulation programs: Uses, problems, procedures, uncertainty, and tools. *ASHRAE Transactions*, *112*(1), 226–240. https://doi.org/Article

Reddy, T. A., & Claridge, D. (2000). Uncertainty of "measured" energy savings from statistical baseline models. *HVAC&R Research*, *6*(1), 3–20. https://doi.org/10.1080/10789669.2000.10391247

Reddy, T., Maor, I., & Panjapornpon, C. (2007a). Calibrating detailed building energy simulation programs with measured data—Part I: General methodology (RP-1051).

HVAC&R Research, *13*(2), 221–241. https://doi.org/10.1080/10789669.2007.10390952

Reddy, T., Maor, I., & Panjapornpon, C. (2007b). Calibrating detailed building energy simulation programs with measured data—Part II: Application to three case study office buildings (RP-1051). *HVAC&R Research*, *13*(2), 243–265. https://doi.org/10.1080/10789669.2007.10390953

Rezaee, R. (2016). *Application of inverse modeling to performance-based architectural design in the early stage*. Georgia Institute of Technology.

Rijal, H. B., Tuohy, P., Humphreys, M. A., Nicol, J. F., Samuel, A., & Clarke, J. (2007). Using results from field surveys to predict the effect of open windows on thermal comfort and energy use in buildings. *Energy and Buildings*, *39*(7), 823–836. https://doi.org/10.1016/j.enbuild.2007.02.003

Roberti, F., Oberegger, U. F., & Gasparella, A. (2015). Calibrating historic building energy models to hourly indoor air and surface temperatures: Methodology and case study. *Energy and Buildings*, *108*, 236–243. https://doi.org/10.1016/j.enbuild.2015.09.010

Robertson, J. J., Polly, B. J., & Collis, J. M. (2015). Reduced-order modeling and simulated annealing optimization for efficient residential building utility bill calibration. *Applied Energy*, *148*, 169–177. https://doi.org/10.1016/j.apenergy.2015.03.049

Robertson, J., Polly, B., & Collis, J. (2013). *Evaluation of automated model calibration techniques for residential building energy simulation*.

Rosen, R. (1991). *Life itself: a comprehensive inquiry into the nature, origin, and fabrication of life*. Columbia University Press.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, *23*(3), 470–472.

Royapoor, M., & Roskilly, T. (2015). Building model calibration using energy and environmental data. *Energy and Buildings*, *94*, 109–120. https://doi.org/10.1016/j.enbuild.2015.02.050

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., … Tarantola, S. (2008). *Global sensitivity analysis: the primer*. Chichester, West Sussex PO19

8SQ, England: John Wiley & Sons Ltd.

Sargent, R. G. (2011). Verification and validation of simulation models. In *Proceedings of the 2011 Winter Simulation Conference* (pp. 2194–2205). https://doi.org/10.1109/WSC.2011.6148117

Schlesinger, S. (1979). Terminology for model credibility. *Simulation*, *32*(3), 103–104. https://doi.org/10.1177/003754977903200304

Srivastav, A., Tewari, A., & Dong, B. (2013). Baseline building energy modeling and localized uncertainty quantification using Gaussian mixture models. *Energy and Buildings*, *65*, 438–447. https://doi.org/10.1016/j.enbuild.2013.05.037

Strachan, P., Hand, J., Svehla, K., Heusler, I., & Kersken, M. (2015). A full-scale empirical validation study applied to thermal simulation programs. *Proceedings of the 14th International Conference of the International Building Performance Simulation Association*, (1997), 2939–2946.

Strachan, P., Svehla, K., Heusler, I., & Kersken, M. (2016). Whole model empirical validation on a full-scale building. *Journal of Building Performance Simulation*, *9*(4), 331–350. https://doi.org/10.1080/19401493.2015.1064480

Strachan P., Monari F., Kersken M., & Heusler I. (2015). IEA Annex 58: Full-scale empirical validation of detailed thermal simulation programs. *Energy Procedia*, *78*, 3288–3293. https://doi.org/10.1016/j.egypro.2015.11.729

Sun, J., & Reddy, T. A. (2006). Calibration of building energy simulation programs using the analytic optimization approach (RP-1051). *HVAC&R Research*, *12*(1), 177–196. https://doi.org/10.1080/10789669.2006.10391173

Sun, K., Hong, T., Taylor-Lange, S. C., & Piette, M. A. (2016). A pattern-based automated approach to building energy model calibration. *Applied Energy*, *165*, 214–224. https://doi.org/10.1016/j.apenergy.2015.12.026

Sun, Y. (2014). *Closing the building energy performance gap by improving our predictions*. Georgia Institute of Technology.

Taheri, M., Tahmasebi, F., & Mahdavi, A. (2012). A case study of optmimization-aided thermal building performance simulation calibration. *Proceedings of the 13th*

International Conference of the International Building Performance Simulation Association, 603–607.

Tahmasebi, F., & Mahdavi, A. (2013). A two-staged simulation model calibration approach to virtual sensors for building performance data. *Proceedings of the 13th International Conference of the International Building Performance Simulation Association*, 608–613.

Tian, W., Wang, Q., Song, J., & Wei, S. (2014). Calibrating dynamic building energy models using regression model and Bayesian analysis in building retrofit projects. *eSim 2014*, (Ashrae).

Tuo, R., & Wu, C. F. J. (2016). A theoretical framework for calibration in computer models: Parametrization, estimation and convergence properties. *SIAM/ASA Journal on Uncertainty Quantification*, *4*(1), 767–795. https://doi.org/10.1137/151005841

Wang, Q. (2016). *Accuracy, validity and relevance of probabilistic building energy models*. Georgia Institute of Technology.

Ward, R., Choudhary, R., Heo, Y., & Aston, J. (2017). A functional principal components model for internal loads in building energy simulation. In *Proceedings of the 15th International Conference of the International Building Performance Simulation Association*. San Francisco: the International Building Performance Simulation Association.

Webster, L., Bradford, J., Sartor, D., Shonder, J., Atkin, E., Dunnivant, S., … Slattery, B. (2015). M&V Guidelines: Measurement and Verification for Performance-Based Contracts - Version 4.0, (November), 1–108.

Westphal, F. S., & Lamberts, R. (2005). Building simulation calibration using sensitivity analysis. In *Proceedings of the 9th International Conference of the International Building Performance Simulation Association* (pp. 1331–1338).

Wikle, C. K., Milliff, R. F., Nychka, D., & Berliner, L. M. (2001). Spatio-temporal hierarchical Bayesian modelling tropical ocean surface winds. *Journal of American Statistical Association*, *96*(454), 382–397.

Xu, B., Fu, L., & Di, H. (2008). Dynamic simulation of space heating systems with radiators controlled by TRVs in buildings. *Energy and Buildings*, *40*(9), 1755–1764. https://doi.org/10.1016/j.enbuild.2008.03.004

Yang, Z., & Becerik-Gerber, B. (2015). A model calibration framework for simultaneous multi-level building energy simulation. *Applied Energy*, *149*, 415–431. https://doi.org/10.1016/j.apenergy.2015.03.048

Yao, W., Chen, X., Luo, W., Van Tooren, M., & Guo, J. (2011). Review of uncertainty-based multidisciplinary design optimization methods for aerospace vehicles. *Progress in Aerospace Sciences*, *47*(6), 450–479. https://doi.org/10.1016/j.paerosci.2011.05.001

Yoon, J., Lee, E. J., & Claridge, D. E. (2003). Calibration procedure for energy performance simulation of a commercial building. *Journal of Solar Energy Engineering*, *125*(3), 251. https://doi.org/10.1115/1.1564076

Zhang, W., & Arhonditsis, G. B. (2008). Predicting the frequency of water quality standard violations using Bayesian calibration of eutrophication models. *Journal of Great Lakes Research*, *34*(4), 698–720. https://doi.org/http://dx.doi.org/10.1016/S0380-1330(08)71612-5